Norges miljø- og biovitenskapelige universitet

Masteroppgave 2021 30 stp
Handelshøyskolen

Et studie av metoder for analyse av boligpriser: Fordeler og ulemper ved bruk av metodene paneldata, tidsserie og prediksjon med regresjonsalgoritmer, for analyse av boligpriser.

An empirical study of methods for analyzing house prices in Oslo

Jacob Haanes Hessen & Øyvind Sveen
Master i økonomi og administrasjon, Business Analytics

Forord

Denne masteroppgaven er det avsluttende, selvstendige arbeidet i masterstudiet økonomi og administrasjon ved Handelshøyskolen på Norges miljø- og biovitenskaplige universitet (NMBU). Oppgaven utgjør 30 studiepoeng innenfor vår mastergrad med spesialisering i Business Analytics.

I løpet av vår tid på NMBU har vi begge fattet interesse for dataanalyse, og har i denne studien utforsket forskjellige metoder for boligprisanalyse. Dette har til dels vært svært krevende, men samtidig utrolig lærerikt. Vi har begge opparbeidet oss interessant og relevant kunnskap og kompetanse vi kommer til å dra nytte av i arbeidslivet.

Vi ønsker å benytte anledningen til å takke alle som har hjulpet oss. Vi ønsker å takke vår veileder Eirik Romstad, som har vært en viktig støttespiller og bidratt med inspirerende ord og konstruktive tilbakemeldinger. Videre ønsker vi å rette en takk til Eiendomsverdi AS for tilgang på datasettet oppgaven bygger analysene på. Avslutningsvis takker vi venner og familie som har holdt ut med oss og bidratt med støtte og motivasjon, takk spesielt til Helen Haanes og Juni Lende for gode tilbakemeldinger.

Ås, august 2021.

Norges miljø- og biovitenskaplige universitet

Jacob Haanes Hessen og Øyvind Sveen

Sammendrag

Denne masteroppgaven undersøker tre forskjellige metoder for boligprisanalyse, der formålet er å utforske metodenes fordeler og ulemper. Da boligprisene i Oslo har steget mye og gjort det vanskelig for førstegangskjøpere å innta markedet, bruker studien typiske førstegangskjøperboliger som avgrensing i analysene. Metodene vi bruker er paneldataanalyse, tidsrekkeanalyse og prediksjon med regresjonsalgoritmer i maskinlæring. Alle analysene bruker et datasett med boligtransaksjoner i Oslo fra 2010-2020, som også inneholder attributter som beskriver hver solgte bolig.

I paneldataanalysen bruker vi en fast effekt-modell, som analyserer månedlig gjennomsnittlig kvadratmeterpris for Oslos 15 bydeler. Med få og nøye utvalgte uavhengige variabler kan vi trekke frem hvordan disse har en fast effekt på gjennomsnittlig kvadratmeterpris. Resultatene viser at økning i boliglånsrenten har en effekt som reduserer kvadratmeterprisen. Videre ser vi at antall kvadratmeter har en negativ effekt, som vil si at man betaler mer per kvadratmeter for mindre boliger. Vi ser også at betydningen av å ha et adskilt soverom i førstegangskjøperboliger, hever kvadratmeterprisen.

I tidsrekkeanalysen tester vi ut en SARIMA-modell, som predikerer gjennomsnittlig kvadratmeterpris i Oslo aggregert på månedsnivå. Her tester vi ut to modeller som predikerer på data vi allerede har. Resultatene her gir lovende resultater som vi videre bruker i en prognosemodell. Prognosen vi lager er på 1,5 år frem i tid, og viser at kvadratmeterprisene kommer til å øke mye. SARIMA-modellen vår gir gode resultater, men har svakheter når det kommer til starttidspunkt for prediksjon. Modellen ser ut til å følge trenden til datapunktene rundt starttidspunktet.

Til sist ser vi på hvordan maskinlæring kan benyttes til å predikere boligpriser ved hjelp av tre-baserte regresjonsalgoritmer. Her ser vi på hvor nærme faktisk kvadratmeterpris modellen estimerer, gitt ett utvalg variabler som beskriver hver boligenhet samt boliglånsrente. Vi tester modellene Decision Trees, Random Forest og Extreme Gradient Boosting. Her oppnår sistnevnte best resultater med en RMSE på 3561,03. Videre benytter vi modellens funksjon som heter «Feature Importance» som et forsøk i å forklare hvordan modellen vekter variablene i prediksjonen.

Hovedkonklusjonen fra denne studien er at metodene har både fordeler og ulemper. Metodene er vanskelige å sammenligne da de har forskjellige formål, men samlet belyser de ulike deler som er nyttige for førstegangskjøpere på jakt etter bolig.

Abstract:

This thesis investigates three different methods for analyzing house prices. The aim is to explore which pros and cons the methods have, and how they contribute to analyzing the real estate market in Oslo. House prices has grown rapidly the ten last years, and smaller dwellings for a typical first-time buyer have become far more expensive than most can afford. Therefore, this thesis uses typical first-time buyer dwellings as a limitation in the analysis. The methods used are fixed effects in a panel data analysis, the univariate time-series model SARIMA, and prediction with the machine learning algorithms Decision Trees, Random Forest and Extreme Gradient Boosting. The three analysis methods use a dataset of transacted properties in Oslo form 2010-2020. The dataset contains attributes which describes each unit sold.

The fixed effects-model in panel data has an advantage in explaining why the average square meter prices varies. The panel data analysis investigates the 15 districts of Oslo as cross-sections and months as the time series. Results here show that an increase in mortgage rate, reduce the average square meter price and vice versa. Further on the analysis shows that smaller dwellings have a higher price per square meter, and that having a bedroom increases the square meter price.

The time-series analysis with SARIMA makes two predictions with in-sample data that produces promising results which we use in an out of-sample forecast 1,5 years ahead in time. The results shows that the SARIMA-model can produce good predictions, but that the model has weaknesses when it comes to the starting point of the prediction. If the datapoints around the starting point has a clear trend the models seem to follow this trend.

In the machine learning analysis, we predict the actual square meter price based on each dwelling´s attributes and the mortgage rate at the selling point. The XGBoost algorithm produced the best prediction with an RMSE of 3561,03. Further on we use XGBoost feature importance in an attempt to understand how the model weights the attributes in its predictions.

The main conclusion drawn from this thesis is that each method has its pros and cons. The panel data analysis is useful in explaining what causes price differences. The time-series analysis is useful in forecasting prices in the future and understanding historical and seasonal variations of the price. And the prediction with machine learning is useful in predicting house prices with high precision but lacks insight in how and why the variables influence the price. The methods cannot be compared directly because of their different perspectives, but all their purposes are all useful for first-time buyers.

Innholdsfortegnelse

Liste med figurer

Liste med tabeller:

1. Innledning

Boligprisene i Oslo har økt kraftig de siste ti årene (Oslo kommune, u.å.). Dette gjør vurderinger som må tas ved et boligkjøp stadig viktigere. Boligprisstatistikk og prisprognoser er vesentlige planleggingsverktøy når en skal investere i bolig (SSB.no, 2019a). Forventninger og endringer i makroøkonomiske forhold som styringsrente og boligrente, o.l, er med på å bestemme hvordan befolkningen velger å bruke penger. Her trekker SSB frem betydningen av statistikker for prisutvikling på boliger. Etter finanskrisen i 2008 anmodet EU om at boligprisindekser skulle være en av hovedindikatorene for å varsle om makroøkonomisk ubalanse (SSB.no, 2017).

Norge har i mange år hatt en boligpolitikk der det å eie sin egen bolig har vært høyt prioritert. Tall fra SSB viser at unge nordmenn flytter tidligere hjemmefra enn i andre land i Europa, og i større grad eier egen bolig. Siden boligprisene har økt de siste ti årene, har det blitt vanskeligere for førstegangskjøpere å komme seg inn på markedet. SSB beskriver at færre unge kjøper bolig og at andelen av unge eiere synker i Oslo (SSB.no, 2019c). En analyse gjort av NBBL viser at single førstegangskjøpere mellom 25-39 år i perioden mellom 2015-2019 kun hadde råd til 5,9% av solgte boliger i Oslo (NBBL, 2020).

Aktører som SSB og Eiendom Norge gir årlig og månedlig ut prisstatistikk og indekser for boligpriser. Disse statistikkene er viktige for befolkningen da nordmenn har den største andelen av formuen sin i bolig (SSB.no, 2019b). Der disse statistikkene ofte beskriver gjennomsnittlige priser for alle boliger, analyserer vi i denne studien typiske boliger for førstegangskjøpere.

Med utgangspunkt i et stort datasett basert på boligtransaksjoner i Oslo og omegn fra 2010-2020, utarbeidet av Eiendomsverdi, så vi som førstegangskjøpere en gylden mulighet til å prøve ut forskjellige metoder for å analysere boligpriser. Først undersøker vi om paneldataanalyse kan gi innsikt i hvordan kvadratmeterprisen har utviklet seg, og om det finnes faste effekter. Videre benytter vi oss av tidsrekkeanalyse med gjennomsnittlig kvadratmeterpris som eneste variabel. Før vi til sist undersøker prediksjon av kvadratmeterpris med bruk av regresjonsalgoritmer i maskinlæring.

1.1 Bakgrunn for studien

1.1.1 Den norske boligmodellen

I Norge eier en stor andel av befolkningen sin egen bolig. Dette har i stor grad sammenheng med boligpolitikken og den norske boligmodellen som strekker seg tilbake til 1920-tallet (Benedictow, A., 2020a). Den norske boligmodellen har som formål at alle som ønsker det skal få eie egen bolig, og være en del av eierlinjen. Tall fra SSB viser at 82,1% av nordmenn eier sin egen bolig, og at mer enn 90 % blir eiere av sin egen bolig i løpet av livet (SSB, u.å.). Den høye andelen av selveiere er et særtrekk for Norge, i forhold til våre naboland. Den høye boligeierandelen er et velferdsgode som innebærer en demokratisering av eierskap i samfunnet. Selveie er også en gunstig spareform, med skattefradrags for gjeldsrenter, lav formueskatt og fraværende gevinstbeskatning ved salg av primærbolig (Eiendom Norge, u.å.a).

Nordmenns viktigste eiendel er altså boligen, som for mange representerer den eneste form for sparing. Tall fra SSB viser at boligformuen var den største formueskomponenten hos nordmenn i 2016 og utgjorde 71% av bruttoformuen (SSB.no, 2018). Nordmenn er også verdensledende når det kommer til oppussing og hjeminnredning målt i utgifter i per innbygger (Forskning.no, 2019). Boligmarkedet og boligpriser har derfor høy interesse hos nordmenn, der avisoverskrifter månedlig bringer nyheter om boligstatistikker og høye prissvingninger. Prisene har i de siste årene gått kraftig opp, noe som er med på å skape usikkerhet. Dette gjelder særlig de som ikke har tilstrekkelig inntekt og tilgang på kapital, men har også betydning for konsumbeslutningen og den personlige økonomien til folk flest. Den høye andelen av boligeiere gjør derfor norsk økonomi følsom for endringer i boligpriser og gjør at boligeierne er en særdeles viktig kraft og pådriver i norsk økonomi.

Som det fremgår ovenfor, er ikke alle omfattet av den norske boligmodellen. Det er viktig at det også opprettholdes et velfungerende utleiemarked. Dette er av betydning for mobiliteten i arbeidsmarkedet, og tar hensyn til folks ulike behov og at folk lever ulike liv (Benedictow, A. et al., 2020b).

1.1.2 Boligprisutvikling i Oslo

I de siste årene har blitt vanskeligere for førstegangskjøpere å komme seg inn på boligmarkedet. Dette skyldes prisstigning kombinert med at boliglånsforskrifter har satt et tak på hvor mye man får låne, ut ifra hvor «rik» man er. I 2015 ble egenkapitalkravet oppjuster fra 10% til 15% som et resultat av høy kredittvekst etter en periode med lave renter etter finanskrisen i 2008. Dette for å redusere kredittrisiko i norsk økonomi (Boliglånsforskriften 1. januar 2020–31. desember, 2020), se vedlegg. Særlig i Oslo har boligkjøperkraften for unge blitt svekket (SSB.no, 2019c).

I 2016 steg boligprisene i Oslo med 26%. Denne utviklingen kom av lave renter grunnet oljeprisfall, samt et begrenset antall boliger på markedet (Rydne, N. og Alsberg, O., 2020). Et resultat av dette var ytterligere boliglånsforskrifter i 2017 med etterfølgende nedgang i boligprisene. Etter denne nedgangen har man sett en økning fra 2018. 2019 bød på moderat økning og koronaåret 2020 hadde rekordmange førstegangskjøpere og økende boligpriser i Oslo. Skulle denne prisveksten fortsette, vil det innebære en stadig vanskeligere situasjon for førstegangskjøperne.

De som da faller utenfor, og ikke får kjøpt bolig vil ikke kunne dra nytte av de norske skattesubsidiene for boligeiere. Forskning viser at land med høyere andel boligeiere har mindre ulikhet og en jevnere fordeling av formuesgoder. Eie av egen bolig viser seg også å ha positiv effekt på boligeierens egen sparing, samt deltakelse i arbeidslivet og produktivitet (Sodini, et al., 2021). SSB har også fremhevet at boligformue er med på å redusere formueforskjellene (SSB.no, 2018).

Det økte presset i Oslos boligmarked danner bakgrunnen for denne studien. Oslo er et populært sted å bo, og de mest sentrale delene er i ferd med å bli for dyre for førstegangskjøpere. Denne kjøpergruppen omfatter unge, til dels nyutdannede mennesker som er av stor betydning for verdiskapning og produktivitet i næringslivet. Vi ser det som nyttig i denne studien å bruke førstegangskjøperen som utgangspunkt for vår analyse. Vi undersøker hvordan ulike metoder kan brukes til boligprisanalyse og sammenligner deres fordeler og ulemper.

Med dagens tilgang på store mengder data og stadig mer oppmerksomhet rundt bruk av maskinlæring, så vi dessuten en gylden mulighet til å tilegne oss relevant kunnskap og bidra til evaluering av boligprisanalyse.

1.2 Boligmarkedet – hva bestemmer boligpriser og hva driver prisene?

I en artikkel i Samfunnsspeilet 2004 (Røed Larsen & Sommervoll, 2004) gjøres en analyse av hva som bestemmer boligpriser. Det konkluderes med at boligpriser settes av en kombinasjon av tilbud og etterspørsel, som påvirkes av realøkonomiens rammer i dag og forventninger til fremtiden. Det understrekes hvordan renter, lønnsnivå og arbeidsledighet har betydning, men også at pessimisme og optimisme har påvirkning på prisutviklingen gjennom forventningskanaler hos banker og husholdninger. Videre fremheves store underliggende endringer i samfunnets struktur. Her nevnes sosioøkonomiske og sosiologiske faktorer som urbanisering, innvandring, familiemønstre, alderssammensetning o.l, som slår ut og påvirker boligmarkedet. Eksempelvis kan prisene på små ettromsleiligheter drives opp av studentinnflytting, småinvestorer som plasserer sparepenger i et marked med høy avkasting, men også av sosiologiske faktorer som høyere frekvens av samlivsbrudd og skilsmisser. Disse eksemplene har en effekt som øker den totale etterspørselen etter mindre leiligheter og kan påvirke prisen.

Ifølge mikroøkonomisk teori bestemmes boligmarkedet i stor grad av tilbud og etterspørsel. Veksten av boligpriser vi nå ser i Oslo er en effekt av at mange vil bo der (økt etterspørsel), samtidig som det er begrensninger i antall nye boliger (begrenset tilbud). Dette presser prisene opp. Eiendomsverdis «sykepleierindeks» illustrerer dette godt. Indeksen viser at single nyutdannede sykepleiere har råd til 3% av boligene i Oslo (Eiendomsverdi, 2019).

For å bremse denne utviklingen sier forskningssjef Erling Røed Larsen på Housing Lab ved OsloMet at «Hvis det bygges mer, blir kampen om boligene svekket.». Her har Housing Lab gjort beregninger som viser at dersom 10 000 boliger bygges i Oslo, og renter og inntekt holdes likt, vil boligprisene falle med 10%, noe som vil medføre at flere førstegangskjøpere for kjøpt bolig (OsloMet, 2020).

1.3 Formålet med studien

Formålet med denne masteroppgaven er å undersøke hvordan forskjellige metoder kan benyttes i boligprisanalyse. Siden metodene vi tester har forskjellige formål, ønsker vi å se hvilke fordeler og ulemper de har for boligprisanalyse. Vi ønsker å presisere at studiet har som formål å vise hvilken nytte data generert gjennom disse metodene kan ha for førstegangskjøpere. Vi har derfor tatt utgangspunkt i enkle modeller og fremgangsmåter.

For å undersøke hva som driver boligprisene, har vi benyttet oss av univariat tidsserie-analyse og paneldata-analyse. Styringsrenten er nå i 2021 særdeles lav, og vil holde seg lav i en periode etter pandemien er over. Selv om det er tegn til bedring i norsk økonomi, er denne bedringen sannsynligvis følsom for nye eksogene sjokk som f.eks. at Covid-19 viruset muterer og kan forlenge restriksjonene i norsk økonomi eller globalt. Boligmarkedet ser imidlertid ikke ut til å være særlig påvirket av denne usikre fremtiden, med rekordmange førstegangskjøpere og høye priser i 2021.

De valgte metodene i dette studiet har forskjellig tilnærming og ulike mål når det kommer til resultater og kan ikke direkte sammenlignes. Vi ønsker imidlertid å besvare i hvilke sammenhenger de forskjellige metodene er hensiktsmessige å bruke, og hva slags og innsikt de kan gi. Problemstillingen i studien vår er derfor:

Hvilke fordeler og ulemper har metodene paneldata, tidsserie og prediksjon med regresjonsalgoritmer, for analyse av boligpriser?

I tillegg til hovedproblemstillingen har vi noen delspørsmål og hypoteser som vi også ønsker å undersøke og besvare i løpet av studie.

Delspørsmål:

- I hvilken grad gir prediksjon med regresjonsalgoritmer innsikt i hva som driver kvadratmeterprisen?
- Hvilke utslagsgivende nøkkelvariabler finnes for boligpriser, og hvordan påvirker disse prisene?
- I hvilken grad gir tidsseriemodeller innsikt i kvadratmeterprisutviklingen?
- Hvilken verdi vil de tre metodevalgene gi samlet?

Hypoteser:

- Prediksjon med regresjonsalgoritmer vil gi presise og gode prediksjoner, men lite innsikt i hva som driver kvadratmeterprisen.
- Sentralitet og boliglånsrente vil i hovedsak være de nøkkelvariablene som har størst påvirkning på kvadratmeterprisutviklingen.
- Tidsseriemodellene vil gi en prognose for kvadratmeterprisen, men ingenting utover det.
- De tre metodene vil gi egne resultater, som samlet vil kunne gi et bedre overblikk over prisutviklingen.

For å kunne besvare problemstillingene vil vi benytte paneldataanalyse for å undersøke hvordan kvadratmeterprisen har utviklet seg på tvers av bydeler over tid. Her bruker vi noen få og nøye utvalgte uavhengige variabler for å se hvordan de påvirker boligprisene.

I tidsserieanalysen bruker vi en SARIMA-modell, hvor formålet er å predikere prisutviklingen med høyest mulig treffsikkerhet, samt gi innsyn i kvadratmeterprisens sesongvariasjoner. Videre bruker vi regresjonsalgoritmer i maskinlæring for å predikere kvadratmeterpris på førstegangskjøperboliger. Maskinlæringsmodellen har som formål å predikere kvadratmeterpris med lavest mulig feilmargin. Dette er en fremgangsmåte som står i sterk kontrast til paneldata der vi antar bydelsmessige forskjeller. SARIMA modeller og maskinlæring skiller seg fra paneldata ved at man her får mindre innsikt i hvilke variabler som påvirker prisen og hvordan disse variablene påvirker boligprisene.

På en måte er univariate tidsrekkemodeller en forløper til maskinlæring ved at disse modellene bruker autokorrelasjonsfunksjoner og partielle autokorrelasjonsfunksjoner til å avdekke systematiske trekk i den avhengige variabelen over tid.

Med økende datakraft og datasett med mange variabler utnytter ikke klassisk tidsrekkeanalysene mulighetene i nye data. En fordel med maskinlæring er dens evne til å se ikke-lineære sammenhenger i data og gi mindre feilmargin, sammenlignet med tradisjonell statistisk modellering (Mullainathan, S., & Spiess, J. ,2017). Ulempen er at man får mindre forståelse for den datagenererende prosessen (i tilfellet vårt boligmarkedet) og hvordan de ulike elementene påvirker prisen. Her bruker vi «Feature Importance» i XGBoost-modellen for å trekke frem hvilke variabler algoritmen vekter som viktigst.

1.5 Avgrensninger

Studiet begrenser seg til førstegangskjøpere i Oslo. Denne kjøpergruppen er mye omtalt i medier og forskning. Blant annet lagde NRK (2021) en kalkulator med fokus på førstegangskjøpere, som beregner hvor og hvilke boliger en har råd til basert på inntekt, gjeld og egenkapital. Førstegangskjøpere er en prissensitiv kjøpergruppe som i hovedsak består av yngre mennesker med relativt lav lønn og lite egenkapital. Denne gruppen frigjør ikke annen bolig ved kjøp, og må derfor låne penger. I rapporten Boliglånsundersøkelse (Finanstilsynet, 2020) kan man se at 49% førstegangskjøpere hadde belåningsgrad på 80-85% og 19% av dem hadde over 85% belåningsgrad. Dette gjør førstegangskjøpere sensitive til endringer i boliglånsrenter.

Som tidligere nevnt har boligprisene hatt en enorm vekst i hovedstaden, der førstegangskjøpere med gjennomsnittlig tilgang på kapital ser ut til å bli drevet ut hvis boligprisutviklingen fortsetter. Førstegangskjøpere betraktes i dette studiet som de som kjøper bolig som er opptil 65kvm. Vi setter en avgrensing på 65kvm fordi boliger som er større enn dette er dyrere enn hva en typisk førstegangskjøper har råd til. Denne avgrensningen gir oss også et solid datagrunnlag med over 100 000 observasjoner.

Videre avgrenser vi oss til å analysere tidsperioden 2010-2020. Denne perioden starter der boligprisveksten så ut til å stabilisere seg i positiv retning etter finanskrisene med negativ

vekst i 2008 og lav vekst i 2009. Siden har prisene økt jevnt, med rekordåret i 2016 og store prisendringer i 2020. Tidsperioden tar også med seg endinger i boliglånsrenter og boliglånsforskrifter. Da denne tidsperioden inneholder flere perioder med unormale prissvingninger, mener vi at dette er et godt utgangspunkt for analyse.

I forhold til metoder vi bruker, vil vi her presiserer avgrensinger. Siden vi foretar flere analyser, har vi holdt analysene relativt simple. I paneldataanalysen bruker vi en fixed effects modell med få forklaringsvariabler. Innenfor tidsserieanalyse finnes det flere modeller og metoder. Vi velger å bruke en SARIMA-modell til å predikere kvadratmeterpris. Valget faller på denne modellen grunnet dens popularitet og simpelhet. Her validerer vi prediksjonsresultatene opp mot de faktiske tallene, for videre og presenterer en prognose med horisont på 1,5 år.

I maskinlæringsanalysen benytter vi tre modeller. Vi velger å holde oss til Decision Trees, Random Forest og Extreme Gradient Boosting. Disse modellene tilhører de populære tre-baserte modellene, og egner seg godt for å predikere ikke-lineære forhold basert på «hvis-så»-regler. Valget av de tre generelle modellen er gjort på grunnlag av erfaringer fra tidligere studier og modellenes popularitet i prediksjonsarbeid. Vårt fokus i analysen vil være modellenes evne til å predikere med høy presisjon. At vi velger å sette søkelys på disse modellene utelukker ikke at det finnes andre modeller, typer og varianter som kan predikere mer presist.

1.6 Struktur

Studiet har som formål å vise fordeler og ulemper ved ulike metoder og tilnærminger for datanalyse av boligpriser, avhengig av hva man ønsker å analysere. Med dataanalyse av boligdata ønsker vi å formidle metodenes styrker og svakheter, samt se hvordan de sammen kan gi et bredere bilde av situasjonen.

I kapittel 2 gjennomgås tidligere litteratur om boligprisanalyser med fokus på maskinlæring. I kapittel 3 gjennomgås datagrunnlaget for de tre analysene. I kapittel 4 presenteres metode.

I kapittel 5 presenteres resultater fra analysene. I kapittel 6 diskuterer vi resultatene, svakheter ved studien og anbefalinger til videre forskning, før vi konkluderer i kapittel 7.

2. Teori og litteratur

2.1 Hedonisk pris modell

I prediksjonsanalysen med maskinlæring er formålet at modellene predikerer kvadratmeterprisen til boliger. Her bruker modellene datavariabler som i hovedsak beskriver interne og eksterne faktorer ved boligene. Vi presenterer derfor teorien rundt en hedonisk pris modell.

Hedoniske prismodeller har som formål å verdsette boliger, med statistiske regresjonsmodeller. Anvendelsen av modellen brukes til å verdsette prisen på boliger med eksterne og interne faktorer som kan beskrive prisen. Der interne faktorer er attributter som beskriver boligen og tomten til boligen. Dette kan for eksempel være størrelsen, antall soverom og boligens tilstandsgrad. De eksterne faktorene er attributter som beskriver boligens beliggenhet, som for eksempel nabolag, nærliggende fasiliteter og kollektive knutepunkter. Her estimeres koeffisienter av hver interne og eksterne faktor til en pris, som samlet utgjør boligens markedspris. Denne metoden brukes av eiendomsmeglere og av Eiendom Norge. Prediksjonsanalysen vår bygges opp på en lignende måte, men vi får ikke innsyn i verdivurdering av hver enkelt faktor.

2.2. Litteraturgjennomgang

Der tidsserie og paneldata allerede er velkjente metoder for boligprisanalyse fokuserer vi i litteraturgjennomgangen på bruk av maskinlæring til boligpris prediksjon. Vi ser at litteraturen rundt maskinlæring for boligpriser er voksende internasjonalt. Prediksjon ved hjelp av maskinlæring har liten utbredelse innen norske studier, men det foreligger konsensus om at maskinlæringsmodeller fanger opp ikke-lineære skjulte sammenhenger mellom variablene i større grad enn tradisjonell lineær regresjon.

I en artikkel fra 2017 undersøker Mullainathan & Spiess hvordan maskinlæring kan brukes i økonometriske sammenhenger. Det poengteres at maskinlæring ikke bare gir økonometri nye verktøy, men at det løser andre problemer i den økonometriske verktøykassen. Maskinlæring har evnen til å se generaliserbare mønstre og oppdage komplekse strukturer

som ikke var spesifisert på forhånd. Den klarer å tilpasse komplekse og veldig fleksible funksjonelle former til dataen, uten «overestimering». Den finner en funksjon som fungerer bra på dataen uten, å ha sett dataen. Videre fremheves hva maskinlæring kan brukes til. Der økonomiske fremgangsmåter ofte er ute etter parameter estimering av parameter β som beskriver sammenhengen mellom y og x, er ikke dette hensikten med maskinlæring. Her poengteres det at maskinlæring ligger i den delen av verktøykassen som er markert med $\hat{Y}$ og ikke $\hat{\beta}$ (Mullainathan, S., & Spiess, J., 2017).

I en studie gjort av Li et al. (2021), bruker de en hedonisk prismodell (HPM) med multippel lineær regresjon (MLR), mot en modell der de bruker Extreme Gradient Boosting (XGB) algoritmen. De bruker XGB ettersomn den hendoniske prismodellen har begrensinger m.h.p. å finne ikke lineære sammenhenger og skille viktigheten av innflytelsesrike faktorer. Studien er gjort for å verdsette innflytelsesrike faktorer som har påvirkning boligmarkedet i byen Shenzhen. Ved å bruke «Feature Importance» i XGB trekker de frem variabler som modellen vekter som viktige. Her finner de at avstand til sentrum, utsikt mot grøntarealer, befolkningstetthet, eiendomsforvaltningsgebyr og økonomisk nivå er de viktigste variablene for prising av boliger i Shenzen.

For å lage prediksjoner av boligprisene startet man tidlig med Artificial Neural Network (ANN). Wilson et al. (2002) så på forskjellen mellom vanlig lineær regresjon på tidsserier og ikke-lineær regresjon ved hjelp av ANN. Forsøket deres viste at ANN kan benyttes til å produsere prognoser med en feilmargin på 3,9 %. Videre viste de at man ved hjelp av en såkalt Gamma test kan tilnærme seg ANN prosessen på en enklere måte. Her benytter man seg av Mean Square Error (MSE) og finner en jevn underliggende modell. Studiet viste at man ved hjelp av ANN kan lage prognoser som er mer nøyaktige.

Lignende studier har blitt gjort med andre former for ANN. Chen et al. (2017) benyttet et Recurrent Neural Network (RNN) kalt Long Short Time Memory (LSTM) til å predikere gjennomsnittlig boligpriser to måneder frem i tid basert på priser i Beijing mellom 2004 og oktober 2016. De benyttet seg av ulike LSTM metoder og sammenlignet mot en tradisjonell Autoregressive Integrated Mowing Average (ARIMA) modell. De fant at LSTM generelt ga bedre prediksjoner enn ARIMA modellen, MSE var 90 % lavere. Problemet i studiet var at datasettet var lite og av begrenset kvalitet. Dette er et problem med ANN da

man er avhengig av tilgang på store mengder data for å lage gode prediksjoner. I tillegg har de ikke fått utnyttet potensialet til det dype neurale nettet da de lagde en stacked LSTM. Dette kan man se på resultatet da de fikk lignende prediksjoner for denne som en vanlig LSTM. Forskjellen mellom stacked og vanlig er antallet skjulte lag. Stacked LSTM har flere skjulte lag. Det er dokumentert i litteraturen at RNN gir bedre og mer presise prediksjoner.

Feng og Jones (2015) tar for seg hvordan beliggenhet og nabolag påvirker boligens verdi. For å se på dette sammenligner de Multilevel Modelling (MLM) og ANN. Disse modellene settes opp mot en Hedonic Price Model (HPM). Funnene fra studiet viser at man får en bedre prediksjon ved å benytte seg av MLM da denne modellen bedre tar høyde for de underliggende faktorene i datasettet.

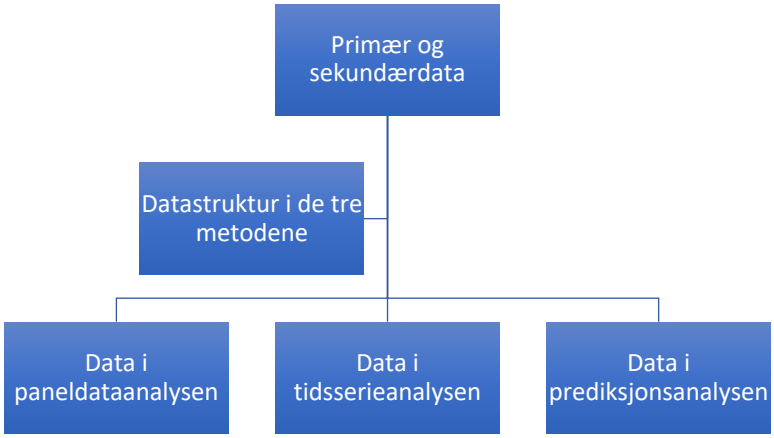
Random forest (RF) har blitt benyttet i stor grad for å predikere prisen på boliger. Wang & Wu (2018) viste at man kan predikere prisen på en bolig med større presisjon ved hjelp av RF enn lineær regresjon. Studiet deres undersøkte få variabler, og man får derfor en relativt lav R^2. Vårt datasett inneholder lignende variabler og studiet er derfor relevant for oss da det viser at RF kan produsere gode prediksjoner.

En samlestudie av Milunovich (2019) tar for seg den australske boligprisindeksen. Studiet forsøker å predikere indeksen 1, 2, 4 og 8 kvartaler frem i tid. Formålet med studiet er å sammenligne flere ulike former for maskinlæring opp mot en random walk benchmark. Studiet konkluderer med at det ved predikering med kort tidshorisont kan lønne seg med en lineær modell, mens det ved lengre tidshorisonter kan lønne seg med ikke-lineær prediksjon. Fordelene med prediksjon blir mindre jo lengre tidshorisonten er, ved 8. kvartal gjør ikke prediksjonen det bedre enn benchmark. En ulempe ved studiet er at de ikke har optimalisert modellene i noen særlig grad.

Vi vet fra tidligere studier at for eksempel neurale nett trenger mye trening for å gjøre det bra (Glorot, X. & Bengio, Y., 2010). Også modeller som Random forest og XGB trenger parameter optimalisering for å få frem potensialet. Dette er forhold som kan forklare hvorfor enklere modeller gjør det bedre i studier da premisset er at det er «first, best forecast» som teller.

3. Data

I dette kapittelet gjennomgår vi primærdatasettet og sekundærdata vi bruker i de tre analysene. Her presiseres de strukturelle forskjellene i hver analysemetode preprossesering og sammenslåinger vi har gjort for at dataen skal passe inn i våre analyser. Til slutt presenteres den avhengige og de uavhengige variablene i de tre analysene. Kapittelet er delt inn i fem deler hvor vi første gjennomgår primær og sekundærdata, før vi gjennomgår data i de tre analysemetodene vi bruker.



Figur 3. 1 Struktur for datakapittel

3.1 Primær og sekundærdata

Primærdatasettet i dette studiet kommer fra Eiendomsverdi AS. Firmaet har Norges største boligdatabase og gir månedlig ut boligprisstatistikk på oppdrag fra Eiendom Norge. Datasettet består av boligtransaksjoner i Oslo og omegn, med en tidshorisont fra 2010 til 2020. Totalt inneholder datasettet 245.437 boligtransaksjoner. Dette er boliger som er formidlet av meglere og annonsert på Finn.no, og er det samme datagrunnlaget Eiendom Norge og Eiendomsverdi bruker i sine boligprisstatistikker (Eiendom Norge, u.å.,b). Datasettet inneholder elleve variabler som beskriver boligenes karakteristikker. Variablene er: *kommune, postnummer, registrering/solgt og tinglysningsdatoer, prisantydning, salgspris, fellesgjeld, størrelse, byggeår, etasje, soverom og eierform.*

Utenom eiendomsdata tar vi med makroøkonomiske sekundærdata som er koblet i samme tidsintervall som eiendomsdataen. Boligmarkedet er sterkt påvirket av hvordan økonomien i landet endres, og endinger i makroøkonomiske faktorer kan ha signifikante korttidseffekter på boligpriser. Vi henter derfor inn variabler for BNP, boliglånsrente, befolkningsvekst, boligprisindeks, konsumprisindeks, og gjeld i husholdninger. Det skal nevnes at vi ikke bruker mesteparten av disse variablene i de endelige analysene, men at de har vært en del av prosessen. Boliglånsrente er i hovedsak den variabelen vi benytter oss av, basert på SSB (2021a) sin gjennomsnittlige rente for pant i bolig. Grunnet forskjeller i tidsrekkene til SSB, har vi kun fått tak i månedlige data for boliglånsrenten fra 2014-2020. Fra 2010-2013 har vi bare kvartalsvise data å basere oss på. Den månedlige boliglånsrenten fra 2010-2013 er derfor basert på interpolering. Gjennomgang av partiell analyse av styringsrente og boliglånsrente, med valg om interpolering ligger i vedlegg.

Videre har vi sammenkoblet datasettet til Bolstad (2009-2021). Dette datasettet kobler alle postnumre med bydeler, og viser også hvilken lengde og breddegrad hvert postnummer har. Denne sammenkoblingen er gjort for å berike datasettet vårt slik at paneldata-analysen kan deles inn i bydeler og, at prediksjon med maskinlæring har flere attributter å predikere kvadratmeterpris ut ifra.

3.1.1 Avgrensinger og endringer i primærdatasettet

Primærdatasettet inneholder alle boligtransaksjoner fra 2010-2020, og inneholder derfor flere boliger som strekker seg utenfor dette studiets avgrensing og omfang. Første steg i endringen er å avgrense datasettet og produsere nye variabler. Til denne studien fjernes derfor alle boligtransaksjoner som er utenfor Oslo. Boliger som er større enn 65 kvadratmeter fjernes også fra datasettet, da boliger over denne størrelsen ikke er typiske boliger for førstegangskjøpere. Etter disse avgrensningene har vi et datasett som inneholder:

Dato for når boligen ble listet, når den ble solgt, når den ble tinglys, postnummer, pris, fellesgjeld, prisantydning, eierform, byggeår, primær rom, bruttoareal, etasje, og antall soverom.

Med utgangspunkt i disse tolv variablene produserer vi tre nye variabler, tid før omsatt, totalpris og kvadratmeterpris.

Tid før omsatt («TOM») lages ved å se på tidspunktet boligen ble listet og tidspunktet den ble solgt, variabelen er da antall dager mellom disse datoene. Totalpris produseres ved å ta summen av fellesgjeld og pris for boliger med andre eierformer enn selveierleiligheter. For selveierleiligheter er totalpris lik salgsprisen. Vi har valgt å regne totalpris på denne måten, da man må ta høyde for fellesgjelden i en lånesøknad når man kjøper en andelsleilighet, noe man ikke trenger dersom man er selveier. Mer om eierform kan leses vedlegg. Kvadratmeterpris lager vi ved å dele totalprisen på boenheten med primærrommets kvadratmetermål. Dette velger vi da primærrom refererer til det arealet av leiligheten som er ment for varig opphold. Denne beregningen er også standard for kvadratmeterpriser hos SSB (2021b).

Etter at de nye variablene er lagt til velger vi å fjerne variablene for listet dato, dato for tinglysning, fellesgjeld og prisantydning før vi eksponerer datasettet til Python for å gjennomføre del to av endringen. I del to kombinerer vi datasettet fra Eiendomsverdi med datasettet til Bolstad (2009-2021), som inneholder alle postnummer og bydelene i Oslo. For å kombinere datasettene benytter vi oss av «merge» funksjonen i Pandas. Dette er en funksjon som muliggjør spleising av data basert på spesifikke detaljer i hvert datasett. I vårt tilfelle er datasettet spleiset på postnummer. Datasettet har etter denne spleisen fått tre nye variabler: bydel, longitude og latitude.

Latitude og longitude fremstiller et geografisk punk i hver bydel og knyttes opp mot postnummeret. Det er flere ulike postnummer og dermed flere geografiske punkter i hver bydel. Disse er dog ikke nøyaktige nok til å gi et godt bilde av hvor boenheten befinner seg, men det er med på å gi maskinlæringsmodellene flere attributter å predikere med.

Videre består datasettet av fire forskjellige eierformer. Disse er selveier, borettslag, aksjeleilighet og obligasjonsleilighet. Av disse fire utgjør transaksjoner av obligasjonsleiligheten kun et tosifret antall. Obligasjonsleiligheter er ikke lenger lov å etablere, og det er lite av disse boligene på markedet (Eiendomsmegler.no, 2019). Vi fjerner derfor disse transaksjonene fra datasettet. De resterende eierformene er beholdt. Videre inneholder datasettet flere transaksjoner med manglende verdier, og ekstremverdier. Disse transaksjonene blir gjort rede for i vedleggskapittel.

Etter korrigeringer og avgrensninger er primærdatasettet forkortet ned til 105.298 boligtransaksjoner, med følgende variabler i tabell 3.1:

Postnummer	P-rom	Boliglånsrente (månedlig)
Bydel	Kvadratmeterpris	Eierform
Etasje	Totalpris	Bredde og lengdegrad
Soverom	TOM	Dato solgt

Tabell 3. 1 Endelige datavariabler

3.2 Datastruktur

De tre metodene vi analyserer krever hver sin datastruktur. I dette delkapittelet presenteres forskjellene i datastruktur og tilpasninger vi gjør for de tre metodene, paneldata, tidsserie-analyse og prediksjon med maskinlæring.

3.2.1 Datastruktur for paneldata-analyse

Paneldata skiller seg fra tidsseriedata og tverrsnittsdata ved at datasettet er strukturert slik at man observerer både tverrsnitt, i, og tidsserie, t, i et og samme datasett (Wooldridge, J., 2019). Datasettet vi har fått tilgang til fra Eiendomsverdi er ikke strukturert som paneldata og dette medfører at vi gjør endringer. Her har vi fordelt alle bydelen i Oslo som hvert sitt tverrsnitt, og bruker en månedlig tidsserie fra 2010 til 2020. Med utgangspunkt i paneldata og dets struktur, er det vanskelig å finne bydelsspesifikke uavhengige variabler som kan forklare prisutvikling. Disse forklarer vi nærmere i delkapittel 3.4. Den avhengige variabelen er her den månedlige gjennomsnittlige kvadratmeterprisen i hver bydel.

I paneldata snakker man om balansert og ubalansert paneldata. Dette beskriver hvorvidt alle tverrsnitts-gruppene har samme antall observasjoner eller ikke. Balansert paneldata vil si at datasettet har like mange observasjoner for alle paneler. Et ubalansert paneldatasett har manglende verdier for én eller flere tidsobservasjoner for én eller flere av gruppene. For å oppnå et balansert datasett fjerner vi en av bydelene, bydel Sentrum som i snitt har veldig få transaksjoner hver måned, og flere måneder uten observasjoner. Ved å fjerne denne bydelen oppnår vi et balansert paneldatasett.

Bydel Sentrum skaper også støy i datasettet ved at få observasjoner skaper et dårlig bilde av gjennomsnittlig kvadratmeterpris, da dette varierer mye fra måned til måned. I prosessen med å omgjøre datasettet til paneldata tester vi også om forskjellene for den gjennomsnittlige kvadratmeterprisen i hver bydel, i, hver måned er signifikant. Testene her er simple lineære regresjoner mellom boligens totalpris og antall kvadratmeter. Vi finner her at alle bydeler utenom bydel Sentrum har nok observasjoner hver måned til å gjenskape et godt bilde av bydelenes typiske kvadratmeterpris-utvikling. Det endelige paneldatasettet har derfor 15 bydels-tverrsnitt med en tidsserie på 132 måneder i hvert tverrsnitt.

3.2.2 Struktur for tidsserie-analyse

Tidsserie-analyse har en datastruktur hvor man undersøker hvordan en avhengig y variabel, varierer over tid, t. Siden tidligere hendelser kan påvirke hva som skjer fremover i tid, er tidsdimensjonen en vesentlig del av tidsserier (Wooldridge, J., 2019). Strukturen for tidsserie er en serie med datapunkter listet i en tidsrekkefølge.
Forskjellen fra paneldatagrunnlaget er at vi her ser på Oslo som en helhet i stedet for per bydel. Dette gjør vi for å kunne si noe om hvordan prisene beveger seg generelt og for å kunne gi en prognose på hvordan de vil bevege seg frem i tid. Vi tar utgangspunkt i det opprinnelige datasettet og trekker ut leilighetene vi ønsker å se nærmere på. Tidsserien vår ser på den gjennomsnittlige kvadratmeterprisen for alle boliger opptil 65 kvm som ble solgt i Oslo mellom 2010 og 2020 og består av 132 månedlige observasjoner.

3.2.3 Datastruktur for prediksjonsanalyse med maskinlæring

I prediksjonsanalysen med maskinlæring skiller strukturen på datasettet seg markant fra hvordan vi har satt opp datasettet i de foregående analysene. Maskinlæringsalgoritmen har som mål å finne mønstre i datasettet, og predikere prisen ut ifra disse mønstrene. Her vekter regresjonsalgoritmene de attributtene som tilhører den enkelte boligen, og finner mønstre som gjør det mulig å vurdere kvadratmeterprisen til boligene.

For å lage en god prediksjonsmodell er det nødvendig med et datasett som inneholder nok observasjoner. Disse observasjonene bør inneholde rikelig med informasjon som beskriver karakteristikker til det en ønsker å predikere.

Lignende studier om maskinlæring og boligprisprediksjon innebærer analyse av langt flere variabler enn det vi bruker, men for denne studien, mener vi at antall observasjoner og variabler i vårt datasett er godt nok til å lage en tilfredsstillende modell.

Sett i sammenheng med tidsserieanalysen og paneldata-analysen, vil maskinlæringsanalysen ta i bruk hele datasettet. Datasettet vi mottok fra Eiendomsverdi inneholder flere av de viktigste og mest relevante variablene for å kunne predikere boligpris. Med over 100.000 boligtransaksjoner har algoritmene rikelig med informasjon å predikere med.

Strukturen på datasettet er satt opp slik at hver bolig har sitt ID nummer, slik at algoritmen kan skille hver leilighet fra hverandre. Dette skiller seg fra de to andre analysene, der tidsaspektet er viktigere. I tillegg skiller datasettet i denne analysen seg fra paneldata ved at bydelene ikke identifiserer som tverrsnitt, men fungerer som en attributt i modellen. Resten av attributtene/uavhengige variablene blir gjennomgått i delkapittel 3.4.2.

For å oppnå et velfungerende datasett har vi foretatt en rekke korrigeringer, slik at maskinlæringsalgoritmene får en enklere jobb. Denne prosessen kalles data preprosessering og er en teknikk som går ut på datarensing og transformering. For at maskinlæringsmodeller skal kunne lese dataen og gi gode resultater er det viktig å gå gjennom rådataen i datasettet. Rådata kan inneholde manglende verdier og avvikende observasjoner som må gjennomgås. Da alle analysene bygger på samme datasettet, er denne prosessen allerede beskrevet i delkapittel 3.1.1, sammen med korreksjoner og endringer som finnes i vedlegg.

3.3 Avhengig variabel

Den avhengige variabelen vi bruker i de tre analyse er kvadratmeterprisen. Grunnet forskjeller i struktur for dataoppsett, ble kvadratmeterpris valgt. Kvadratmeterpris beskriver hvor mye man må betale per kvadratmeter, og er den prisvariabelen som brukes i prognoser for boligprisindekser og prisstatistikk. Kvadratmeterpris gjør det lettere å se på den faktiske kostnaden på boligmarkedet, på samme måte som kilopris gjør det enklere å sammenligne prisen på lignende produkter i matbutikken.

3.3.1 Paneldata

I paneldata-analysen er den månedlige gjennomsnittlige kvadratmeterprisen for hver bydel den avhengige variabelen. Dette gir oss muligheten til å analysere bydelenes forskjeller i kvadratmeterpris over tid. Målet i denne analysen er å se hvordan de uavhengige variablene, som presenteres i neste delkapittel, påvirker kvadratmeterprisen på tvers av bydeler og over tid.

3.3.2 Tidsserie

I tidsserieanalysen er den månedlige gjennomsnittlige kvadratmeterprisen for Oslo den avhengige variabelen. I denne analysen er det den eneste variabelen som analyseres sammen med de månedlige tidsintervallene. Målet i denne analysen er å se hvordan den sesongbaserte tidsseriemodellen SARIMA, predikerer fremtidig gjennomsnittlig kvadratmeterpris basert på sesongvariasjoner og tidligere strukturer i tidsserien.

3.3.3 Maskinlæring

I prediksjonsanalysen med maskinlæring er den faktiske kvadratmeterprisen for hver enkelt bolig den avhengige variabelen. Sett i sammenheng med lignende studier og fremgangsmåter for boligprisprediksjon, er den avhengige variabelen ofte boligens totalpris. For at studiet skal ha en sammenhengende variabel som kan analyseres, har vi også her valgt kvadratmeterpris. Målet med denne analysen er å se hvilken regresjonsalgoritme som predikere faktisk kvadratmeterpris med høyest mulig presisjon.

3.4. Uavhengige variabler

3.4.1 Uavhengige variabler i paneldata-analysen

Siden vi benytter oss av gjennomsnittlige kvadratmeterpris i paneldataanalysen, medfører dette at vi lager lignende uavhengige variabler. Med denne metoden er det begrenset med uavhengige variabler vi kan ta i bruk. Dette er fordi vi ser på snittet av mange boliger. Variablene vi lager er derfor også snittverdier.

Snitt av p-rom

For hver bydel og hver månedlig tidsperiode regner vi ut en snittverdi for antall kvadratmeter boligene har. Snitt av p-rom har liten variasjon da boligene vi analyserer har relativt lav spredning i antall kvadratmeter. Uansett vil dette gi en indikasjon på verdien av en ekstra kvadratmeter for boliger til førstegangskjøpere.

Snitt av soverom

For hver bydel og hver månedlig tidsperiode regner vi ut en snittverdi for antall soverom boligene har. Snitt av soverom har liten variasjon da boligene vi analyserer er små, der mange ikke har adskilte soverom. Uansett vil dette gi en indikasjon på verdien av et soverom for boligene vi ser på.

Månedlig boliglånsrente

Månedlig boliglånsrente er med for å fremheve sammenhengen den har med kvadratmeterprisen. Denne variabelen varierer kun over tid, t, og er den eneste uavhengige variabelen i paneldata-analysen som ikke også varierer på tvers av bydel.

AR(12) ledd

For å kunne si noe om sesongvariasjonen i bydelsanalysen med paneldata, tar vi med et autoregressivt ledd. AR(12)-leddet er en autoregressiv funksjon av kvadratmeterprisen som vi lagger 12 måneder tilbake i tid. På den måten får vi en variabel som kan si noe om sammenhengen mellom kvadratmeterprisen i tidsperiode og bydel $Y_{i,t}$ og $Y_{i,t-12}$, der Y er kvadratmeterprisen.

3.4.2 Uavhengige variabler i prediksjonsanalysen med maskinlæring

Postnummer

Postnummer gir en geografisk beskrivelse av hvor boligen befinner seg. Lokasjon av bolig har stor betydning for boligkjøpere. Sentrumsnære, populære bydeler og strøk, tilgang til natur og knutepunkter i kollektivtrafikk er med på å drive prisene opp.

Denne variabelen er derfor viktig i prediksjon, da vi også vet at sentralitet er en betydelig faktor som driver prisene opp (OsloMet, 2020).

Eierform

Variabelen eierform er en kategorisk variabel som beskriver hvorvidt boligen har eierformen aksjebolig, selveier eller borettslag. Ved kjøp av en bolig i borettslag kjøper man en andel av et borettslag med borett i en bolig. Borettslaget er et selskap som eies av de som bor der. Bygging av borettslag finansieres delvis med et felleslån med sikkerhet i borettslagets eiendom og innskudd fra boligkjøperen. Innskuddet for å kjøpe boligen, pluss andelen av boligens fellesgjeld utgjør den totale prisen for boligen (USBL.no, u.å). Det samme prinsippet gjelder for aksjeboliger. Her kjøper man ikke boligen, men en aksje i boligaksjeselskapet. Denne aksjen gir deg leierett i den bestemte boligen du var ute etter (Krogsveen, 2021). Også for aksjeboliger kan det være tilknyttet fellesgjeld. Vi legger derfor sammen salgsprisen og fellesgjelden for å få boligens totale verdi. Med en selveierbolig derimot, kjøper man en sameieandel som gir kjøper enerett til bruk av en bestemt bolig. For selveierboliger er prisen det vi benevner som totalpris. Fellesgjelden er ofte lavere som selveierboliger, og blir ikke tatt med i beregning av totalpris.

I forhold til boligenes verdi kan eierform være med på å heve eller senke prisen. Mindre leiligheter som kan leies ut, har blitt et populært investeringsobjekt i Oslo, blant de med nok kapital. Med selveierboliger har kjøper rett til å leie ut boligen fra første dag. Her kan borettslag ha begrensinger for utleie, det samme gjelder for aksjeleiligheter. I boligselskap kan også den økonomiske driften spille en rolle for boligens verdi. Alt i alt er det flere aspekter ved eierform som kan være med på å påvirke boligens kvadratmeterpris.

P-rom

P-rom står for primærrom og beskriver hvor mange kvadratmeter boligen er. Her inneholdt primærdatasettet også verdier for bruksareal (BRA), men siden kvadratmeterpris blir beregnet ut ifra P-rom (SSB.no, 2021b) har vi valgt å ikke bruke BRA.

Etasje

Etasje er en numerisk verdi beskriver hvilken etasje boligen ligger i.

Antall soverom

Antall soverom er en numerisk verdi og beskriver hvor mange soverom boligen har. Her har flere av boligene ingen soverom. Årsaken til dette er studiens omfang og avgrensing på maks 65 kvadratmeter store boliger. Flere av de mindre boligene i Oslo har ikke adskilte soverom, men f.eks alkover.

Bydel

Bydel er en kategorisk variabel som gir en geografisk beskrivelse av hvor boligen befinner seg i Oslo.

Lat og Long

Lat og long står for latitude og longditude, på engelsk og beskriver boligens bredde og lengdegrad. Disse variablene beskriver boligens geografiske lokasjon ut ifra hvilket postnummer de har. Selv om disse variablene viser lokasjonen til postnummeret, kan regresjonsalgoritmene fange mønstre som priser boligene basert på hvilken himmelretning de ligger i.

TOM

TOM som står for «tid før omsatt» beskriver hvor lenge boligen har ligget ute for salg på markedet.

Månedlig boliglånsrente

Månedlig boligrente er en numerisk variabel som beskriver den gjennomsnittlige renten for pant i bolig. Denne renten er som tidligere beskrevet sterkt korrelert med styringsrenten, og derfor en viktig brikke i konsumbeslutningen til befolkningen. Ettersom konsum er med på å bestemme etterspørselen i boligmarkedet, kan boliglånsrenten være en viktig variabel i prediksjonsmodellen.

Måned solgt og år solgt

Måned solgt og år solgt er variabler som baseres på hvilken dato boligen ble solgt. Ettersom kvadratmeterprisen har hatt en tydelig trend fra 2010-2020, er disse variablene med på å gi algoritmen en indikasjon på hvordan pristrenden ser ut i det gitte tidsrommet boligen ble solgt.

Flere av variablene vi bruker i denne analysen er kategoriske variabler. For at algoritmene skal kunne tyde disse, har vi i Python brukt Scikit-learn sin preprosseseringsmetode som heter «Lable Encoder». Denne metoden gir de kategoriske variablene en numerisk verdi.

4. Metode

4.1 Introduksjon

I dette kapittelet presenterer vi den empiriske fremgangsmåten og metoden vi bruker i de tre analysene. Først presenteres fremgangsmåte og metode, samt problemer som oppstår ved bruk av paneldata. Deretter forklarer vi det teoretiske rammeverket som ligger bak modellen vi bruker. Deretter presentere vi metode og utfordringer ved bruk av tidsserie-analyse. Til slutt presenterer vi metode, fremgangsmåte og utfordringer for prediksjon med maskinlæring. Resultater kommer i kapittel 5. Maskinlæringsanalysen og tidsserieanalysen vil i dette studiet foregå ved hjelp av Python og i paneldataanalysen benytter vi Stata, som er en velkjent programvare innen økonometri.

4.1.1 Forskjeller mellom maskinlæring og statistiske metoder

Hovedforskjellen mellom statistiske modeller og maskinlæringsmodeller er at maskinlæring har som mål å lage de mest nøyaktige prediksjonene, hvor statiske modeller er laget for å si noe om forholdet mellom variablene og signifikansen av disse forholdene. Med maskinlæring deler man opp datasettet i et treningssett og et testsett. Treningssettet brukes til å trene maskinlæringsalgoritmen, hvor den lærer seg komplekse sammenhenger mellom den avhengige og de uavhengige variablene. Testsettet brukes så til å validere treffsikkerheten til modellen. I testsettet vet ikke modellen hva den avhengige variabelen er, og den bruker det den lærte fra treningssettet til å predikere den avhengige variabelen. Statistisk modellering derimot, analyserer regresjonsparameterne med konfidensintervall, signifikanstester og andre tester som kan si noe om modellens legitimitet (Stewart, M., 2019).

4.2 Paneldatametode og teoretisk modell

Paneldata-modellering handler i stor grad om å vise til den sannsynlige avhengigheten på tvers av dataobservasjoner innen samme gruppe.

Paneldata skiller seg fra tidsseriemodeller ved at paneldatamodeller tillater heterogenitet på tvers av grupper, og dermed fremhever individuell-spesifikke effekter. Paneldatamodeller har altså en struktur som kan adressere heterogenitet på tvers av observasjonene.

Siden vi har strukturert dataene våre i et panel, har vi med to forskjellige dimensjoner. Tidsdimensjonen blir omtalt som t, og bydelsindividene blir omtalt som i. Med utgangspunkt i en enkel paneldatamodell med en forklaringsvariabel, kan det presenteres slik:

$$y_{it} = \beta_1 x_{it} + \alpha_i + u_{it} \quad (4.1)$$

Likningen ovenfor representerer en heterogen modell, dette er fordi konstantleddet, α_i, er individspesifikt. Den gjennomsnittlige kvadratmeterprisen, y, i hver bydel, i, i de forskjellige tidsperiodene, t, varierer hver for seg og over tid, og derfor er modellen vår bydelspesifikk.

Paneldata observasjonen Y_{it}er gjennomsnittlig kvadratmeterpris observert for bydelene $i = 1,\ldots,15$ over alle tidsperioder $t = 1,\ldots,132$. Paneldatasettet består av 15 bydeler som fordeler seg over 132 tidsperioder, som beskriver antall måneder fra 1/1/2010-1/12/2020. $\beta_1 x_{it}$ beskriver betakoeffisienten til forklaringsvariabel x, for i og t. Variabelen α_i er et tidskonstant feilledd som kun varierer på tvers av bydel, i. Dette feilleddet beskriver den uobserverte effekten i hver bydel. Variabelen u_{it} er et feilledd som varierer over i og t. Dette feilleddet beskriver den idiosynkratiske feilen, som er uobserverte faktorer som har en innvirkning på den avhengige variabelen.

Observerte variabler er boliglånsrente, bydelssnitt for primærrom og snitt for soverom og et AR(12) ledd av den avhengige variabelen. Uobserverte variabler vil være andre variabler som ikke er observert eller med i datasettet, men som har en forklaringskraft for kvadratmeterprisen sin utvikling i hver av bydelene. En uobservert faktor vil være hva som skiller det gjennomsnittlige prisnivået mellom hver av de 15 bydelene. Et problem i paneldata er at utelatte og uobserverte variabler vil kunne gjøre modellen forventningsforskjøvet og ukorrekt (Wooldridge, J., 2019). Dette er fordi utelatte variabler fanges opp i forklaringsvariabelen X_{it}, og vil korrelere med α_i. Dette bryter med en av OLS-forutsetningene, og resulterer i å benytte mer avanserte estimeringsmetoder som kalles fixed effects (FE) og random-effects (RE).

4.2.1 Paneldata-modeller

I paneldatamodeller er det stor sannsynlighet for at uobserverte faktorer konsekvent påvirker utfallet, altså y variabelen. For å ta høyde for dette tar vi i bruk estimeringsmetodene Fixed effects og Random-effects.

4.2.2 Fixed effects estimator

Fixed effects estimatoren, heretter FE, er den mest brukte metoden for å fange opp forskjeller mellom et tverrsnitt av observasjoner over tid. Formålet med denne metoden er å fjerne den uobserverte effekten, α_i, fordi uobserverte effekter kan være vilkårlig relatert til de observerte forklaringsvariablene (Wooldridge, J., 2019). FE estimering gjøres ved å transformere dataen. Her kalkuleres snittverdien for en variabel over tid for hvert individ, og deretter trekkes dette snittet fra alle observerte verdier for et gitt individ, dette gjøres så for alle individene. Denne metoden bearbeider de uobserverte faktorene ved at den fjerner de komponentene som er konstant over tid. Som antakelse ville det vært alle de uobserverte variablene (Pedace, R., 2013). Denne metoden gjøres enkelt i programvaren Stata, og vil her forklares med likninger.

Modellen spesifiseres først slik:

$$(Y_{it} - \bar{Y}_i) = \beta_1(X_{it} - \bar{X}_\iota) + \beta_2(\omega_{it} - \bar{\omega}_i) + (\varepsilon_{it} - \bar{\varepsilon}_\iota) \quad (4.2)$$

$$\tilde{Y}_{it} = \beta_1 \tilde{X}_{it} + \tilde{\varepsilon}_{it} \quad (4.3)$$

Hvor

$$\bar{Y}_\iota = \frac{\sum_{t=1}^{T} Y_{it}}{T} \quad (4.4)$$

$$\bar{X}_\iota = \frac{\sum_{t=1}^{T} X_{it}}{T} \quad (4.5)$$

$$\bar{\omega}_i = \frac{\sum_{t=1}^{T} \omega_{it}}{T} \quad (4.6)$$

Her er $\beta_1$gitt som FE estimator, også kalt innenfor estimator. Den uobserverte variabelen ω har blitt transformet bort siden verdiene er antatt konstante over tid.

I likning (4.3) har variablene blitt transformert til deres tids-nedjusterte versjon, også kalt innenfor transformasjon.

FE transformerer dataen slik at en ser på avvik fra det det individspesifikke gjennomsnittet. Ved å gjøre dette isoleres variasjonen i hver enkelt bydel, i, når man estimerer. Denne metoden fjerner den individuelle komponenten til restleddet. Dette gjør at tidskonstant og ikke observerte forskjeller mellom bydelene transformeres bort. FE undersøker om konstantleddet i endrer seg i bydelene eller mellom tidsperiodene.

4.2.3 Random-Effects Estimator

Som vi har vært gjennom tidligere i kapittelet kan man benytte seg av FE modeller for å estimere effekter man ikke kan observere. En annen metode for å estimere slike effekter er med tilfeldig heterogenitet modellen, kalt «random effects estimator», heretter RE. Der tilfeldig heterogenitet estimeres med modellen:

$$Y_{it} = \beta_0 + \beta_1 X_{it} + v_{it} \tag{4.7}$$

der $v_{it} = w_i + \varepsilon_{it}$. (4.8)

Fordelen med tilfeldig heterogenitet modellen over faste effekter modellen er at man kan estimere regresjonsparameterne mer effektivt. Når man studerer modellen for tilfeldig heterogenitet ser man at den skiller seg fra FE-modellen ved at den ikke estimerer de faste effektene for hvert individ, i vårt tilfelle bydelene. Dette gjør at man har færre estimerte parametere, flere frihetsgrader og mindre standardfeil (Pedace, R., 2013).

RE-modellen benyttes om man tror at den uobserverte effekten ikke er korrelert med de uavhengige variablene. Dersom den er det, vil man ha et forventningsforskjøvet estimat. Selv om disse forutsetningene er til stede, betyr ikke dette at effektene er konstante eller like for alle observasjonene. Det betyr derimot at alle effektene er tilfeldige og uavhengige av de observerte variablene. For å se på hvilken modell man bør velge kan man benytte seg av en test utviklet av Hausman i 1968, som vil gjennomgås i neste avsnitt.

4.2.4 Hausman test for fixed eller random-effects

Hausman-testen er en spesifikasjonstest som tester hvorvidt en skal bruke FE eller RE modellen til å estimere paneldatasettet. RE gir mer effektive estimater enn FE, men hvis individuelle faste effekter er korrelert med en eller flere uavhengige variabler vil modellen være forventningsskjev. Da vil FE være foretrukket. Hausman-testen sjekker ut Random-effects antakelsene og hjelper med valget mellom de to metodene. Testen undersøker forskjeller i estimerte parametere, og resultatene i testen bestemmer om FE og RE estimatene er signifikant forskjellige. Nullhypotesen til testen sier at hvis antakelsene til RE holder, vil RE resultere i like estimerte parametere som FE. Man vil da velge RE fordi den er bedre i form av effektivitet og har mindre standardfeil. Hvis antakelsene til RE ikke holder, er de estimerte parameterne signifikant forskjellige og RE vil være forventingsforskjøvet. Dette resulterer i den alternative hypotesen, som vil si FE estimatene er konsekvente (Pedace, R., 2013).

4.2.5 Den økonometriske paneldata-modellen

Gjennom den økonometriske modellen ønsker vi å forklare drivere for den gjennomsnittlige kvadratmeterprisen. Da Oslo-boligmarkedet er ujevnt fordelt i forhold til priser på boliger, vil panel-data analysen gi et bilde av hvordan den gjennomsnittlige kvadratmeterprisen har utviklet seg på tvers av bydeler. Modellene er bygd opp med følgende forklaringsvariabler: Månedlig boliglånsrente, snitt av primærrom for hver bydel hver måned og bydelenes årlige befolkningsutvikling. Vi har i tillegg lagt til en AR-komponent som variabel. Denne har som mål å forklare prisutviklingen ut ifra forrige års kvadratmeterpris.

Forklaringsvariablene gir et bilde av hvordan de påvirker kvadratmeterprisen. Da disse variablene på ingen måte fanger opp alt som påvirker prisutviklingen og forskjeller mellom bydelene, vil det være mange uobserverte effekter. Det er disse effektene vi ønsker at modellene i paneldata-analysen skal fange opp og forklare. Her tester FE modeller som blir presentert i kapittel 5.

Vår modell:

$$P_{it} = \beta_0 + \beta_1 Prom_{it} + \beta_2 Rente_t + \beta_3 Soverom_{it} + \beta_4 AR12_{it} + a_i + u_{it}$$

P_{it} = Gjennomsnittlig kvadratmeterpris for boliger i bydel i, og måned t

$Prom_{it}$= Gjennomsnittlig primærrom (kvadratmeter) for boligene i bydel i, og måned t

$Rente_t$ = Boliglånsrente i måned t, for alle bydeler

$Soverom,_{it}$ = Snitt av antall soverom for boliger i bydel i, og måned t

$AR12_{it}$= AR(12) ledd for gjennomsnittlig kvadratmeterpris for bydel i, og måned t

u_{it}= Feilleddet til modellen i bydel i, og måned t

a_i= Den uobserverte effekten

FE = Fixed effect = $a_i + u_{it}$

4.3 Metode for tidsserieanalyse med SARIMA

De lineære statistiske modellene for tidsserier er relatert til lineær regresjon, men forklarer også sammenhengen som oppstår mellom datapunkter i samme tidsserie. I motsetning til standardmetoder brukt på tverrsnittsdata, antas det at hvert datapunkt er uavhengig av de andre datapunktene i utvalget (Nielsen, A., 2019). I tidsseriedata har datapunkter som ligger nær hverandre i tid en tendens til å være sterkt korrelert med hverandre, og det er her tidsseriemodellene skiller seg fra lineær regresjon. Siden tidsserien legger til en gitt struktur som dataen følger, må man med tidsseriedata sørge for at tidsserien er stasjonær. Når observasjonene i tidsserier er stasjonære, er trender differensiert bort og det er tilnærmet lik varians gjennom hele tidsserien. Sammendragsstatistikk som gjennomsnitt og variansen av observasjonene i tidsserien vil da være konsekvente over tid. En tidsserie som har en trend, sesongvariasjoner eller andre strukturer som avhenger av tiden, er hva man kaller en

ikke-stasjonær tidsserie. Her vil variansen og gjennomsnitt av observasjonene endre seg over tid og derfor gi hint av de konseptene den eventuelle modellen kan ha mål i å fange.

Innenfor tidsserieanalyse finnes det flere modeller og metoder. I dette studiet bruker vi en enkel univariat modell.

Denne heter SARIMA og er en metode for å modellere matematiske modeller som brukes til prognoser. Metoden bruker tidligere verdier fra tidsseriedata og et feilledd til å estimere fremtidige verdier.

4.3.1 Test for stasjonæritet

Vi tester derfor for stasjonæritet med ADF-test. Augmented Dickey-Fuller testen er en "unit root test" som sier hvor sterkt en tidsserie er definert av en trend. Testen bruker en auto-regressiv modell og optimerer et informasjonskriterium over flere forskjellige laggede verdier. Testen forteller oss at vi ikke kan forkaste nullhypotesen om at tidsserien er stasjonær. Metoden vi bruker for å oppnå stasjonæritet, er å differensiere tidsserien til vi kan forkaste nullhypotesen i ADF-testen. Resultatene her kommer i kapittel 5.

4.3.2 Autoregressive (AR) modeller

AR modeller bygger på intuisjonen om at fortiden forutsier fremtiden, der verdien på et tidspunkt (t) er en funksjon av tidsseriens verdier på tidligere tidspunkter. Den enkleste AR modellen AR(1) beskriver følgende:

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + u_t \tag{4.8}$$

Verdien av serien på tidspunkt t er en funksjon av konstantleddet β_0, dens verdi i det forrige tidsstempelet multiplisert med en annen konstant $\beta_1 Y_{t-1}$ og et feilledd som også varierer med tid u_t. Feilleddet antas å ha konstant varians og et gjennomsnitt på 0. AR(1) beskriver et tilbakeblikk på ett lag, den ser altså tilbake til tiden med ett lag.

4.3.3 Moving Average (MA)

En MA-modell bygger på et sett med prosesser der verdien på hvert punkt i tid er en funksjon av nærliggende tidligere verdier av feilleddet, der disse er uavhengig av de andre. En MA-modell ligner en AR-modell unntatt at begrepene i den lineære likningen refererer til nåværende og tidligere feilledd istedenfor nåværende og tidligere verdier av selve prosessen. En MA-modell med ordre q kan beskrives slik:

$$Y_t = \mu + \varepsilon_t + \vartheta 1\varepsilon_{t-1} + \vartheta 2\varepsilon_{t-2} \ldots + \vartheta q\varepsilon_{t-q} \quad (4.9)$$

Der tidsserien Y_t er en lineær kombinasjon av feilleddet ε_t, eller «white noise» som det kalles på engelsk. Dette beskriver antakelsen at hvert element i en tidsserie er tilfeldig fra en populasjon med konstant varians og et snitt på 0. Her er μ konstantleddet og ϑ er modellens parameter som kalles theta.

4.3.4 Seasonal autoregressive integrated moving average (SARIMA)

En SARIMA modell kombinerer AR(p) og MA(q) modellene og gjenkjenner at den samme tidsserien både kan underliggende AR og MA-modell dynamikker. Dette tilsvarer en ARMA modell, men en ARIMA modell vil her inneholde differensiering som er en måte å fjerne trender og gjengi tidsserien stasjonært. Differensiering (d) går ut på å konvertere en tidsserie med verdier til en tidsserie med endringer i verdier over tid. Så verdien av den differensierte serien i tid t er verdien av tid t minus verdien av tid t-1, dette kan gjøres på forskjellige lag intervaller. Forskjellen mellom en ARMA og ARIMA modell er at ARIMA inneholder begrepet (i) integrert, som refererer til antall ganger tidsserien differensieres for å produsere stasjonaritet.

Et problem med ARIMA-modellen er at den ikke støtter sesongvariasjon i dataen. Av den grunn bruker vi SARIMA-modellen som legger til sesongkomponenter og fire nye parametere. Der ARIMA modellen bruker parameterne (p,d,q), legger SARIMA til (P,D,Q)s. Der:

p = Antall lag av avhengig variabel (AR)

d = antall differensieringer

q = Antall lag av feilleddet (MA)

P, D, Q = Refererer til de sesongbaserte AR, D og MA parameterne og s, som refererer til antall perioder per sesong.

For å finne hvilken ordre parameterene skal ha, bruker vi et ACF-plot (Autocorrelated fuction) og PACF-plot (Partial autocorrelated fucntion). ACF er en autokorrelasjonsfunksjon og PACF en partiell autokorrelasjonsfunksjon. ACF plottet viser hvor godt nåværende verdier i tidsserien er relatert til tidligere verdier.

PACF plottet viser residualene, altså det som er igjen etter at man fjerner effekten som allerede er forklart med tidligere verdier. PACF plottet viser derfor gjemt informasjon som ligger i feilleddet som kan bli modellert med neste lag.

Her har SARIMA syv parameter, der det krevers ekspertise og nøye analyse for at modellen treffer. Vi bruker her et Grid-Search i Python som metode. Grid-search modellen har som mål å finne optimale parameterverdier. Basert på analyse av ACF og PACF plots gir vi modellen et søksområde av parameterverdier. Her søker vi etter den kombinasjonen med parametere som gir lavest AIC-score. Akaikes informasjonskriterium (AIC) estimerer den relative verdien en modell mister.

4.4 Fremgangsmåte for prediksjon med maskinlæring

Med de dataene vi har samlet og gjennomgått i kapittel 3, ønsker vi å predikere kvadratmeterprisene ved hjelp av maskinlæring. I en prediksjonsmodell er formålet til algoritmen å finne og modellere underliggende forhold mellom en målvariabel, i vårt tilfelle kvadratmeterprisene for en leilighet, og tilhørende attributter. Eller som Kuhn og Johnson (2013:2) skriver, prediksjonsmodellering er «… the process of developing a mathematical tool or model that generates an accurate prediction.”. I kapittel 5 tester vi alle modellene for å finne den med høyest prediksjonsnøyaktighet. Nøyaktigheten måler vi ved hjelp av RMSE.

Formålet med denne analysen er å få en så god prediksjon som mulig fremfor å kommentere på sammenhengen mellom de ulike variablene i datasettet. Uansett forsøker vi å finne hvorfor den beste modellen predikerer som den gjør.

Vi har valgt å benytte oss av Python til maskinlæring da man har tilgang til mange gode kilder og utvalget av biblioteker er svært stort. I Python er et bibliotek en samling verktøy, og av disse kommer vi primært til å benytte oss av Scikit-Learn. Fordelen med dette biblioteket er at man finner verktøy for å dele datasettet i trening og testsett, man har maskinlæringsmodellene vi ønsker å benytte og man kan med enkelhet finne informasjon og hjelp til å benytte biblioteket. Dette gjør at man lett kan komme i gang med maskinlæring selv med liten kunnskap om temaet fra før.

Figur 4.1 viser en standard fremgangsmåte vi bruker. Denne er ment som illustrasjon i hva som blir gjennomgått videre i kapittelet.

Figur 4. 1 Fremgangsmåte i maskinlæring

4.4.1. Overvåket og ikke-overvåket læring

Maskinlæring er et vidt begrep og omfatter en rekke teknikker og algoritmer. Vi benytter oss av det man kaller «svak kunstigintelligens» i vår analyse. Dette betyr at maskinen er programmert til å gjøre en spesifikk oppgave. Maskinlæring kan primært gjøres på to ulike måter når det kommer til læringsprosessen; overvåket og ikke-overvåket læring (Hastie, T., Tibshirani, R. & Friedman, J., 2008). Når man benytter seg av overvåket læring ønsker man at modellen skal se etter en sammenheng i datasettet og komme frem til en prediksjon av en

variabel basert på flere input-variabler. I vårt tilfelle er dette en prediksjon av kvadratmeterprisen.

Ved ikke-overvåket læring benytter man seg av en annen strategi. Her handler det i større grad om at datasettet skal finne mønstre og sammenhenger på egenhånd. Ikke-overvåket læring foregår ved at man ikke har en klart definert målvariabel og ingen tydelig oppgave som skal løses. Algoritmens jobb blir å se de underliggende sammenhengene i datasettet og kategorisere disse.

4.4.2. Trening og testdata

For å kunne lage prediksjonsmodeller med tilfredsstillende grad av presisjon er det viktig med gode forberedelser av datasettet. Det er spesielt viktig at data som benyttes til prediksjonen ikke også er inkludert i datasettet som benyttes til trening. Dersom man benytter seg av data i prediksjonen som også er inkludert i treningssettet vil dette redusere presisjon av prediksjonen.

Treningsdata er den delen av det opprinnelige datasettet man benytter til treningen av maskinlæringsalgoritmen. Denne delen legger grunnlaget for hvordan man justerer parametere for å få en så nøyaktig prediksjon som mulig. En stor andel av datasettet benyttes til trening, dette gjør man fordi man ønsker stor variasjon og et godt grunnlag når man skal gjøre prediksjonene. Testdata er den delen av datasettet som er igjen etter at man har tatt ut det som er nødvendig til trening. Dette er nødvendig for å sikre at modellen produserer troverdige prediksjoner i møte med «nye» data. For å få en god fordeling av trenings- og testdata er det viktig å ha i bakhodet mengden data man har tilgjengelig for å unngå problemer med enten overestimering eller underestimering. Hvordan datasettet deles opp avhenger også av hvilken type data man ønsker å predikere på og størrelsen på datasettet. Vi benytter en 80:20 fordeling på vårt datasett da utvalget for trening vil være svært godt og vi fremdeles har ett stort datagrunnlag igjen for test av modellen. Trening og testdata blir i litteraturen også referert til som henholdsvis «in sample» og «out of sample».

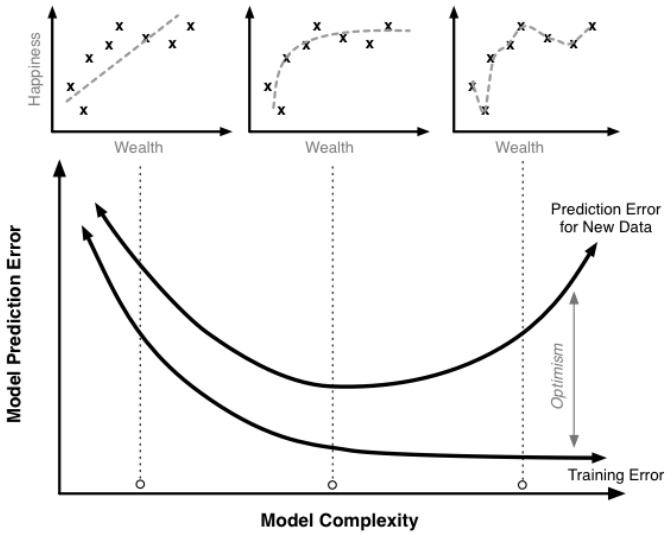
4.4.3 Avveining mellom forventningsforskyving og varians

Maskinlæring har blitt ett populært analyseverktøy da man kan finne sammenhenger i store datasett uten å måtte programmere innstillingene for å finne disse. I boligmarkedet kan man finne store mengder data ved å bruke script som tråler for eksempel finn.no. Det er store mengder statistikk på ssb.no, norges-bank.no og lignende sider. I tillegg til de offentlig tilgjengelige kildene sitter Eiendomsverdi og andre private aktører på data som ikke er tilgjengelige uten et abonnement. Hvor nyttig all informasjonen er i en analysesammenheng, avhenger av hvordan man ønsker å benytte seg av data man samler inn i til prediksjonsformål.

For å være nyttige må modellene man lager inkludere variabler som er med på å klassifisere eller inneholder data som gjør det mulig å predikere et utfall eller et resultat. Dersom man øker antallet variabler i modellen øker man også kompleksiteten. Problemet med økt kompleksitet er at modellen vil slite med å skille mellom signalet og støyen i datasettet. Dette viser Fortman-Roe (2012a) på to ulike måter. Først presterer han en formel for faktisk prediksjonsfeil:

$$True\ Prediction\ Error = Training\ Error + f(Model\ Complexity) \tag{4.10}$$

Formelen viser at den faktiske prediksjonsfeilen kan deles i to ledd, treningsavvik og modellkompleksitet. Med dette viser Fortmann-Roe at ved å øke modellkompleksiteten vil treningsavviket i modellen bli redusert. Dette vil skje uavhengig om variabelen som introduseres og øker kompleksiteten er relevant eller ikke. Videre viser han dette i en illustrasjon av prediksjonsavvik og kompleksitet som kan ses i spredningsplottet i figur 4.2.



Figur 4. 2 Trening, optimisme og faktisk prediksjonsfeil

I dette spredningsplottet vises en modellering av lykke opp mot rikdom. Fortmann-Roe (ibid.) øker kompleksiteten til modellen ved å inkludere flere variabler for velstand.

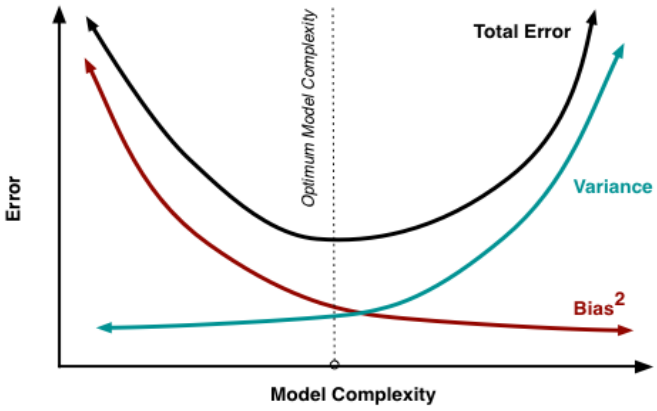
Figur 4.2 viser hvordan prediksjonsavviket for trenings- og testdata endrer seg etter hvert som kompleksiteten i modellen øker. Til venstre i modellen ser vi at man har store avvik for prediksjonen i begge datasettene. Etter hvert som kompleksiteten øker, reduseres avviket. Til høyre i modellen øker avviket for testdatasettet mens man stadig ser en reduksjon i avviket for treningsdata. Den optimale tilpasningen for modellen finner man i området der økt kompleksitet i modellen ikke øker avviket for prediksjonen av testdatasettet. Når man benytter seg av maskinlæring for å predikere ved hjelp av overvåket læring er en av de store utfordringene overestimering.

Overestimering omfatter alle modellene med høyere kompleksitet enn det optimale området. Dette er et problem fordi det representerer en modelltilpasning som tar for stor høyde for kompleksiteten i treningsdatasettet og prøver å tilpasse seg slik at den eksakt passer med alle punktene i treningsdatasettet. Stilt ovenfor testdatasettet vil modellen ikke klare å predikere på et tilfredsstillende nivå da datapunktene her ikke vil være like. Man kan si at modellen ikke generaliserer i stor nok grad og dermed er for optimistisk. For å finne den riktige kombinasjonen av variabler må man stille dette opp med hvor stort datasett man har. Dersom man har mange observasjoner over en lengre tidsperiode kan man ofte se at noen variabler faller ut etter hvert som man går bakover i

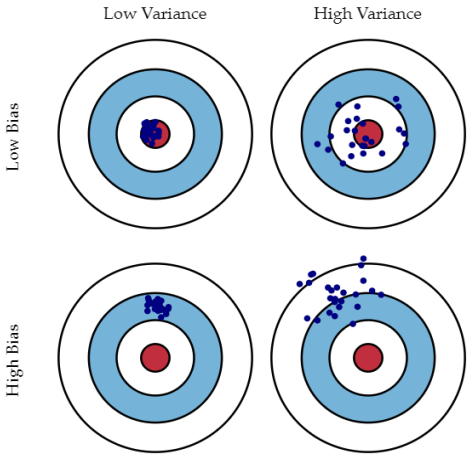
tid. Dette skyldes at kvaliteten på data og tilgjengeligheten har blitt større i senere tid. Dette fører oss videre til det man i litteraturen beskriver som bias-variance trade-off, (James, et al. 2017).

Man kan tenke på bias som avstanden mellom predikert verdi og faktisk verdi. Varians kan man tenke på som ulikheten mellom prediksjoner på det samme datapunktet.

Fortmann-Roe (2012b) illustrer hvordan bias og varians påvirker prediksjonen i figur 4.4 og hvordan bias og varians kan benyttes til å finne den optimale modell kompleksiteten i figur 4.3.



Figur 4. 4 Grafisk illustrasjon av hvordan bias og varians påvirker prediksjon. (Fortmann, R. S., 2012b)



Figur 4. 3 Grafisk illustrasjon av hvordan bias og varians bidrar til å finne skjæringspunktet for optimal kompleksitet. (Fortmann, R. S., 2012b)

4.4.4. Kryssvalidering

Når man benytter seg av overvåket læring i prediksjonssammenheng er det viktig å huske på at man ved å dele datasettet inn i trening- og testdata også reduserer mengden data man kan benytte til å utvikle modellen. Videre kan det oppstå et problem dersom man har en

kompleks modell, hvor maskinen tillegger mye verdi til variabler som ikke er relevante. Dette problemet kalles som nevnt i forrige delkapittel, overestimering.

Modellen tar for mye hensyn til støy i datasettet og det igjen medfører at modellen produserer svært gode prediksjoner på treningsdata, men dårlige prediksjoner på testdata. Validering av modellen er derfor svært viktig for å se til at prediksjonene er robuste og ikke har for høy skjevhet (bias) eller for stor varians. Man kan se om modellen overestimerer ved at prediksjonene blir bedre på treningsdatasettet enn det blir når man skal teste modellen på data som ikke ble brukt til å utvikle modellen.

Når modellene er ferdig med treningssettene kan man ikke gå ut ifra at modellen vil ha den sammetreffsikkerheten og variansen på testdatasettet. For å overkomme dette benytter vi 10-Fold kryssvalidering. Denne metoden er enkel og den mest bruke metoden til å estimere prediksjonsfeil. Her deles datasettet opp tilfeldig i ti deler, hvor modellen trener seg på ni deler og tester på en del. Denne prosessen gjøres ti ganger, slik at alle delene i datasettet til slutt har blitt brukt til testing. Her lagres testresultatene fra hver «Fold», slik at modellen vet hvor godt den treffer på hver av de ti delene. Dette gir et snitt av «out of sample error», som gir et mer presist estimat av modellens sanne prediksjonsfeil for «out of sample» data (Hastie et al., 2009).

4.4.5 Forarbeid før prediksjon

Maskinlæring handler om å gi en mengde med data til en algoritme, slik at maskinen kan lære av dette og lage en modell som predikerer den valgte variabelen. For at man skal kunne lage gode prediksjoner, må datasettet ha høy kvalitet. I kapittel 3 beskrev vi hvordan vi ryddet opp i datasettet og la til flere variabler. For å se om disse variablene egner seg til prediksjon er det viktig å foreta undersøkelser av datasettet.

Disse undersøkelsene blir benevnt som EDA og står for «Exploratory data analysis». EDA gir oss innsyn datasettet, noe som er viktig før man går i gang med prediksjon. Det er i denne undersøkelsen vi finner de unormale datavariaiablene og manglende verdiene vi beskrev i kapittel 3. Videre bruker vi en korrelasjonsmatrise for å undersøke hvordan variablene i datasettet korrelerer. Et vesentlig problem med mange variabler er multikollinearitet. Dette

oppstår når to eller flere variabler har høy korrelasjon med hverandre. Dette kan gjøre det vanskelig å skille den individuelle effekten disse variablene har på kvadratmeterprisen.

Her foretar vi en VIF-test som står for «variance inflation factor», hvor en VIF-verdi på over 5 eller 10 er problematisk (James, et.al, 2013). Dette resulterer i at variablene BNP, gjeld i husholdninger, KPI, BPI, BRA fjernes.

4.4.6 Algoritmer

Mange av de mest populære algoritmene for overvåket læring faller inn under tre kategorier. Lineære modeller forsøker å finne den best passende linjen gjennom datapunktene ved hjelp av en enkel formel.

Tre-baserte modeller benytter seg av en rekke «hvis-så» regler for å predikere basert på ett eller flere valgtrær. Kunstige neurale nettverk er den siste kategorien og her prøver man enkelt forklart å etterligne hvordan neuroner i den menneskelige hjernen tolker informasjon og løser problemer.

Lineære modeller egner seg godt til å predikere enkle sammenhenger, men når det kommer til mer kompleks data der sammenhengende er ikke-lineære, slik som med boligpriser er ikke dette modeller som egner seg. Det motsatte gjelder for kunstige neurale nettverk. Dette er modeller som er svært komplekse og kan gi gode prediksjoner. Slike modeller blir ofte benyttet til å tolke bilder og tekst. Ulempen med disse modellene er at de krever mye datakraft for å trene, samtidig som de ikke er like enkle å tolke.

Vi har derfor valgt å benytte oss av den andre gruppen med modeller. Fordelen med tre-baserte modeller er at de er gode til å tolke ikke-lineære sammenhenger samtidig som de ikke krever alt for mye å trene. Ulempen er at de kan ha en tendens til å overestimere. Vi vil nå presentere de ulike modellene vi velger å benytte oss av for å predikere kvadratmeterprisene.

4.4.6.1 Decision Tree

Decision Trees, “valgtrær” på norsk, er en teknikk der man starter med en rotnode. Fra denne rotnoden deler man valgtreet opp i ulike grener som er knyttes til flere valgpunkter.

Disse valgpunktene vil nå kalles noder. Den endelige prediksjonen kalles terminalnoden, og lages når noden ikke kan generer flere nye grener (Theobald, 2017). Valgtrær er en overvåket algoritme som fungerer både på klassifisering og regresjons problemer. I dette studiet velger vi å benytte valgtrær fremfor multippel lineær regresjon (MLR). Denne metoden vil bedre fange opp både lineære og ikke-lineære forhold mellom målvariabelen og inputvariablene, noe som kan være vanskelig med MLR fordi da må angi riktig funksjonell form.

Målet er å lage en modell som predikerer verdien til målvariabelen, basert på enkle valgregler utledet fra datasettet. Man kan derfor se på valgtreet som en konstant tilnærming av målvariabelen, her kvadratmeterprisen.

Vi har benyttet scikit learn når vi har bygget modellen og deres valgtre-algoritme baserer seg på CART (Classification and Regression Trees) (Pedregosa et al. 2011).

For at valgtreet skal kunne lage gode prediksjoner er det viktig å sette grenser for hvor stort valgtreet kan vokse. I vårt valgtre har vi satt maks dybde til 10 for å hindre at treet overestimerer. For å sikre at modellen ikke skal overestimere videre benytter vi også 10 folds kryssvalidering.

4.4.6.3 Random Forest

Valgtre modellen er noe begrenset ved at man der kun benytter seg av ett tre. Vi tar hensyn til at valgtre modellen ofte kan overestimere ved at vi benytter oss av kryssvalidering.

Selv med tiltak vil en slik modell være utsatt for feil knyttet til forventningsrettfeil på grunn av for mange restriksjoner på målfunksjonen og variansfeil som en følge av at små endringer i treningsdatasettet fører til store endringer i prediksjonen (Glen, 2019).

Vi benytter oss derfor av Random Forest, som er en samling av valgtrær, og på grunn av dette er den mindre sårbar for å overestimering sammenliknet med et enkelt valgtre. Den potensielle overestimeringen reduserer ved at modellen gror et stort antall mindre valgtrær, en “skog”, basert på “bootstrapped training samples” (Hastie, T., Tibshirani, R. & Friedman, J., 2008, s.249).

Bootstrapped aggregation blir ofte forkortet til «bagging» og er viktig for hvordan Random Forest modellen blir bygget opp.

Bagging gjøres ved at man tar tilfeldige utvalg fra datasettet før man så gjør regresjoner på hvert av disse utvalgene. Videre kombinerer man alle resultatene for å finne en representativ prediksjon (Breiman, 1996). Siden man trekker ut ett likt antall tilfeldig variabler fra populasjonen, er det forskjellige variabler som danner grunnlaget for hvert utvalg. De ulike variablene danner grunnlaget for splittene for nye grener i hvert tre, og man gjør så regresjoner på hvert utvalg. Når dette er gjort føres utvalget tilbake til populasjonen og man trekker så ett nytt utvalg og repeterer. For hvert utvalg man tar finner man den mest egnede variabelen og splitte treet videre på.

Man repeterer prosessen tilstrekkelig antall ganger, og man vil da sitte igjen med ett stort antall noder. Ved å benytte seg av et vektet gjennomsnitt av prediksjonen fra de ulike nodene vil man finne den endelige modellen for prediksjonen.

Problemet med en ren bagging fremgangsmåte er at sterke variabler kan havne i flere av nodene. Med sterke variabler mener vi variabler som er viktige for at modellen skal gi god prediksjon. Dersom disse variablene havner i flere av nodene vil man få noder som korrelerer. Et eksempel for oss er om vi hadde benyttet både p-rom og BRA som korrelerer sterkt i prediksjonen. Nodene vi hadde fått da, ville hatt overrepresentasjon av disse variablene som korrelerer med hverandre og som har en individuell sammenheng med målvariabelen. Vi ville dermed kunne fått en modell som overestimerte.

Random Forest modellen bygger på det samme prinsippet og bygger flere valgtrær med ulike noder trukket tilfeldig med tilbakelegging fra det opprinnelige datasettet. Random Forest skiller seg fra «bagging» ved at man også legger inn et tilfeldig valg i hver node, slik at variabelen som skal bestemme splitten i valgtreet også er tilfeldig. Det randomiserte valget gjør at man forhindrer at de samme variablene blir benyttet i hver node. Slik forhindrer man at modellen overestimerer og man får valgtrær som i større grad representerer hvordan ulike variabler påvirker boligprisen. Videre for å sikre mer robuste resultater benytter vi oss av et «grid search» søk der vi legger inn en rekke ulike parametere. Formålet med søket er å

finne den kombinasjonen av parametere som gir lavest «Mean Average Error». Denne kombinasjonen benyttes så videre i den endelige modellen.

Random Forest modellen er en anvendelig modell som egner seg godt til flere ulike prediksjonsformål. Vi benytter modellen fordi den håndterer ikke-lineære sammenhenger på en god måte. Samtidig er metoden god fordi den ved hjelp av randomiseringen i de ulike leddene forhindrer overestimering og varians i modellen. Ulempen med metoden er at innsikten i hva algoritmen gjør er begrenset. Dette gjør det utfordrende å poengtere hvorfor modellen predikerer godt eller dårlig.

4.4.6.4 Extreme Gradient Boosting

"Extreme Gradient Boosting" heretter "XGBoost" er en algoritme som i likhet med Random forest er basert på valgtrær.

Forskjellen er at algoritmen i stedet for «bagging» benytter seg av «Gradient Boosting». Gradient boosting ble først introdusert i Greedy Function Approximation: A Gradient Boosting Machine (Friedman, 2001).

Gradient boosting kan benyttes til å løse både klassifikasjon og regresjonsproblemer og er en samlingsalgoritme i likhet med Random forest. Fremgangsmåten for algoritmen bygger på at svake modeller bygger på hverandre for til slutt å danne en robust modell. Dette gjøres ved at de svake modellene i hvert steg danner grunnlaget for den neste modellen. For hvert steg sitter man igjen med ett feilledd som ikke lar seg forklare av den svake modellen. I det etterfølgende steget er dette feilleddet fokuset. Slik fortsetter modellen å forbedre seg ved å ta hensyn til feilene i foregående ledd. Til sist samles alle leddene og danner den endelige modellen. XGBoost er en forlengelse av Gradient boosting algoritmen. For det første muliggjør algoritmen benyttelse av alle kjerner på datamaskinen, slik at tiden benyttet på beregningen forkortes. For det andre inneholder algoritmen instrumenter for å sørge for at trærne ikke vokser seg større enn en gitt størrelse. Dette medfører at deler av treet som ikke tilfører modellen økt prediksjonsevne fjernes slik at sjansen for forventningsskjevheten og variansen reduseres og modellen blir mer robust. For det tredje inneholder algoritmen et ledd for regularisering. Dette er en regresjon som krymper koeffisientestimatene mot null og

på den måten sørger for at modellen ikke blir for fleksibel og med det overestimerer (Jorly, 2020).

Også for denne modellen benytter vi oss av kryssvalidering og en «grid search» for hyperparametere for å finne de beste parameterne for modellen. I tillegg foretar vi en undersøkelse av de ulike variablene for å se hvilke som gir mest informasjon til algoritmen og dermed forbedrer prediksjonen. For å gjøre dette benytter vi oss av den innebygde funksjonen i XGBoost biblioteket kalt Feature Importance. Feature importance kan kalkulereres på tre ulike måter, og vi velger å benytte oss av metoden kalt «Gain». Resultatet fra denne funksjonen vil da kunne si noe om hvilke variabler som i gjennomsnitt bidrar til å forbedre prestasjonen ved å være grunnlaget for en node. Denne forbedringen måles i en F-score, som står for feature-score.

4.4.7 Ytelse og måltall for vurdering av presisjon

Når man jobber med maskinlæring, er man avhengige av å ha verktøy for å skille mellom gode og dårlige modeller. Slike verktøy finner man i form av beregninger for flere interessante aspekter ved algoritmen. Vi benytter oss av tre mye brukte beregningsmetoder for å analysere prestasjonen til modellene våre, MSE, RMSE og MAE. Både MSE og RMSE ser på residualene til prediksjonen, der residualene er forskjellen mellom de faktiske verdiene og de predikerte verdiene.

Mean Squared Error (MSE) er et måltall for snittet av den kvadrerte differansen mellom predikerte og faktiske verdier. Formålet med beregningen er å finne variansen i residualene.

Root Mean Squared Error (RMSE) er en beregning av standardavviket til residualene. Formålet med denne beregningen er å få et mål på spredningen av residualene og med det indikere hvor konsentrer prediksjonene er rundt de faktiske verdiene (Holmes, 2000). RMSE kan uttrykkes som kvadratroten av de gjennomsnittlige kvadrerte residualene (Formel 4.2)

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}} \tag{4.2}$$

Siden RMSE kan uttrykkes slik, medfører dette at man kan lese av resultatet med lik skala og enhet som den avhengige variabelen. For oss betyr dette at dersom RMSE er 1500, så vil vår

predikerte kvadratmeterpris i snitt ligge 1500 kr fra den faktiske kvadratmeterprisen. En viktig egenskap til RMSE er at den straffer store avvik hardt. Dette er en følge av at residualene blir kvadrert, der store avvik straffes hardere sammenlignet med små avvik. Som et evalueringsverktøy er dette bra, da man ønsker modeller som har små avvik i prediksjonene.

I tillegg til RMSE benytter vi oss av Mean Absolute Error (MAE) for å måle prediksjonsevnen. MAE skiller seg noe fra RMSE, da man her ser på den gjennomsnittlige absolutte størrelsen på residualene. Med den samme notasjonen som for RMSE kan MAE utrykkes som følgende:

$$MAE = (\sum_{i=1}^{n} |\hat{y}_i - y_i|)/n \tag{4.3}$$

På samme måte som med RMSE kan man benytte samme skala og enhet som vår avhengige variabel. En MAE på 1000 betyr derfor at prediksjonen i snitt predikerer kvadratmeterprisen 1000 kroner feil per kvadratmeter. Det som skiller RMSE og MAE er hvordan residualene behandles. Siden MAE ikke kvadrer residualene, skiller den heller ikke på små og store avvik i prediksjonen. Dette medfører at MAE egner seg noe mindre som ett enkeltstående mål på prediksjonen. Vi benytte MAE sammen med RMSE fordi vi da kan få en dypere forståelse av resultatene. RMSE vil aldri kunne gi et lavere estimat enn det MAE gjør. Dette utnytter vi ved å se på forskjellen mellom to modeller som har relativt lik MAE, men forskjeller når det kommer til RMSE. Man vil da kunne se at modellen med høyest RMSE har relativt større avvik i prediksjonen enn den andre.

I tillegg benytter vi oss av R2 for å evaluere hvor godt modellene klarer å fange opp variansen. R2 er et praktisk mål i den innledende programmeringen av modellene siden en høy R2 score på treningsdata-prediksjonen og en lav score på testdata-prediksjonen indikerer at algoritmen overestimerer modellen.

5. Resultater

I dette kapittelet presenterer vi resultatene fra hver analyse. Først presenterer vi deskriptiv statistikk samt grafer og plots for å visualisere dataen og testresultater. Deretter går vi nærmere inn på resultatene av hver analyse.

5.1 Paneldataresultater

5.1.1 Sammendragsstatistikk

Vi spesifiserer modellen vår:

$$P_{i,t} = \beta_0 + \beta_1 Boliglånsrente_t + \beta_2 Snitt\ soverom_{i,t} + \beta_3 Snitt\ Prom_{i,t} + \beta_4 P_{i,t-12} + a_i + u_{it}$$

Der $P_{i,t}$ beskriver gjennomsnittlig kvadratmeterpris for bydel, i, og tidsvariabel månedår, t.

Variable		Mean	Std. Dev.	Min	Max	Observations	
BydelID	overall	8	4.321694	1	15	N =	1800
	between		4.472136	1	15	n =	15
	within		0	8	8	T =	120
månedår	overall	671.5	34.64944	612	731	N =	1800
	between		0	671.5	671.5	n =	15
	within		34.64944	612	731	T =	120
Gjenno~s	overall	59976.54	16623.61	27034.84	109463.2	N =	1800
	between		11100.53	42280.01	79968.78	n =	15
	within		12699.34	34513.56	89658.69	T =	120
Gje~erom	overall	1.142456	.173392	.25	1.75	N =	1800
	between		.123684	.8440473	1.39825	n =	15
	within		.1256143	.2926419	1.800998	T =	120
Gje~prom	overall	49.45627	3.746401	36.5	62.25	N =	1800
	between		2.984469	44.51647	55.399	n =	15
	within		2.391159	37.80626	62.91101	T =	120
Boligl~e	overall	3.08975	.752062	1.76	4.18	N =	1800
	between		0	3.08975	3.08975	n =	15
	within		.752062	1.76	4.18	T =	120
AR12	overall	3942.997	5726.28	-21541.28	28533.75	N =	1800
	between		801.5306	2583.116	5000.566	n =	15
	within		5673.652	-21061.18	29013.84	T =	120

Tabell 5. 1 Sammendragsstatistikk for paneldata

I tabell 5.1 ser vi sammendragsstatistikken for variablene vi bruker i paneldataanalysen. Først ser vi bydelID som beskriver en numerisk verdi fra 1-15 for hver bydel. Deretter kommer månedår, som beskriver tidsperioden. Videre kommer gjennomsnitt av kvadratmeterpris, gjennomsnitt av antall soverom, gjennomsnitt av p-rom, boliglånsrenten og AR12-leddet.

Hver av variablene har en overall, between og within variasjon:

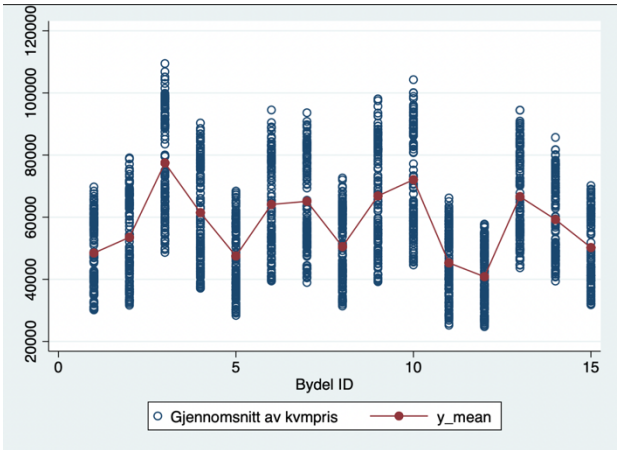
- Within: Hvor mye variasjon det er i samme bydel over tid.
- Between: Hvor mye variasjon det er mellom bydelene.
- Overall: Hvor mye det varierer i forhold til det totale gjennomsnittet.

Fra sammendragsstatistikken ser vi at boliglånsrenten er den eneste variabelen uten «Between» variasjon. Dette kommer av at renten er lik for alle bydeler *i*, og kun varierer over tid *t.* De resterende variablene er bydelsspesifikke og endrer seg med tid.

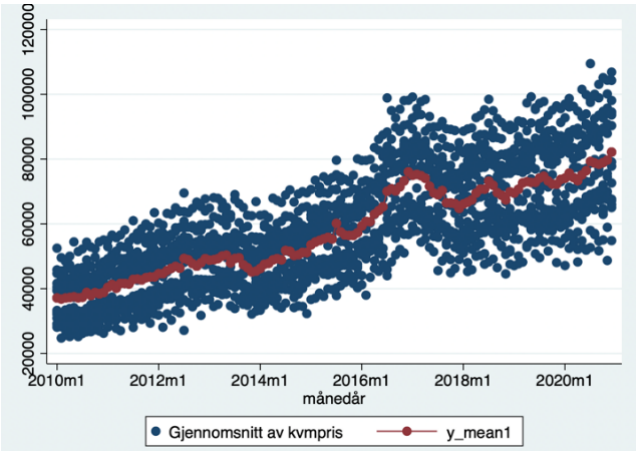
Sammendragsstatistikken viser at den gjennomsnittlige kvadratmeterprisen har høyest variasjon i forhold til det totale gjennomsnittet. Deretter varierer kvadratmeterprisen mer innen hver enkelt bydel, enn hva den gjør på tvers av bydelene. «Within» variasjonen beskriver utviklingen i kvadratmeterprisen fra 2010-2020 innen hver bydel. Dette forteller oss at kvadratmeterprisen sin trend fra 2010-2020 har mer variasjon, enn variasjonen mellom bydelene. Videre ser vi antall observasjoner som er 1800, som beskriver 120 tidsperioder for de 15 bydelene.

5.1.2 Visualiseringer for paneldataanalysen

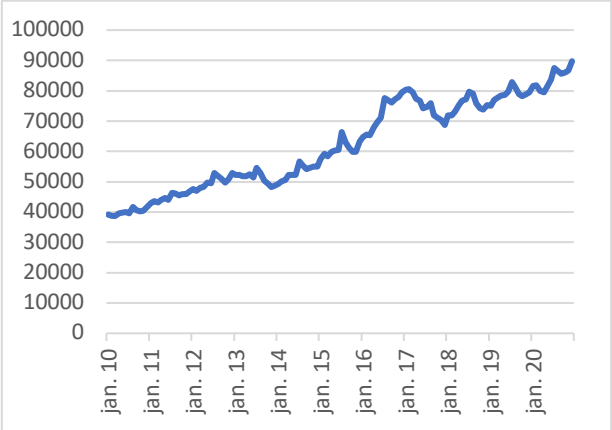
For å få et overblikk av hva vi analyserer, presenterer vi fire figurer av den gjennomsnittlige kvadratmeterprisen i hver av de 15 bydelene fra 2010-2020.



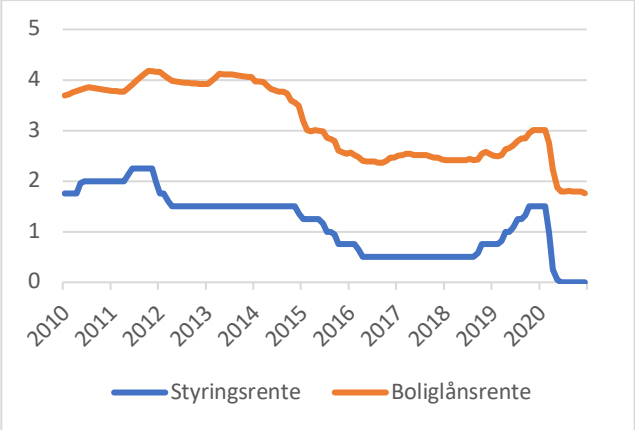
Figur 5. 1 Heterogenitet på tvers av bydelene



Figur 5. 2 Heterogenitet over måneder og år



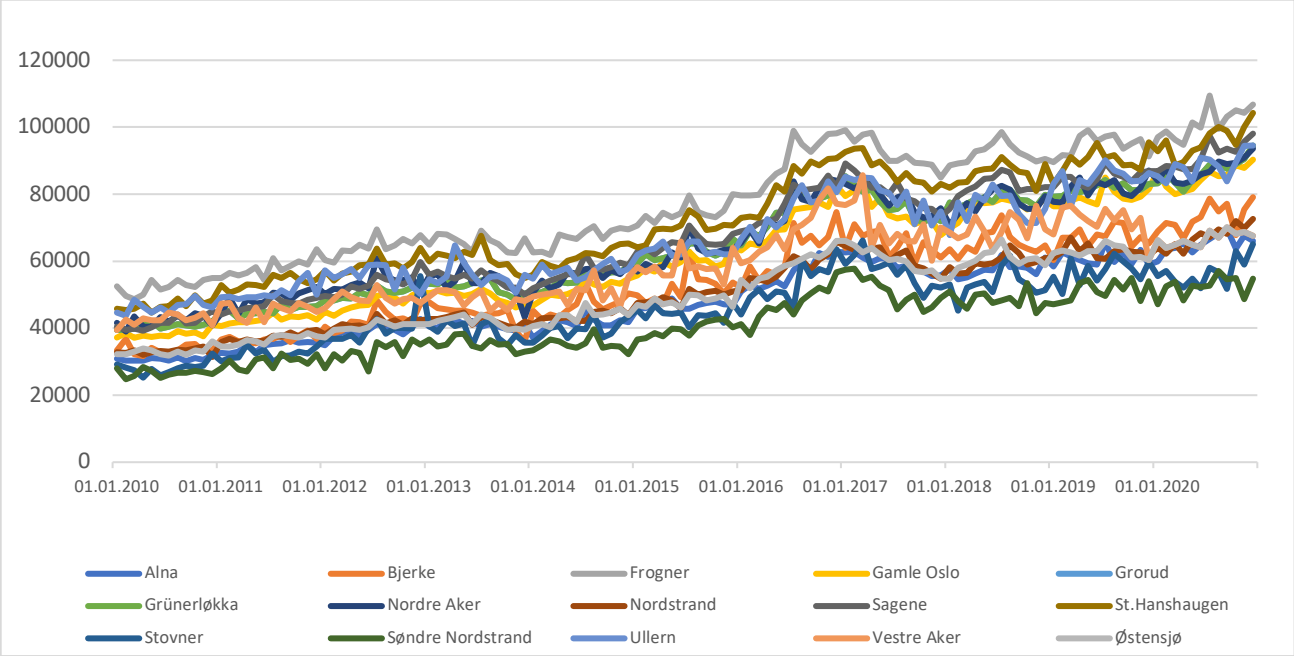
Figur 5. 4 Gjennomsnittlig kvadratmeterprisutvikling



Figur 5. 3 Renteutvikling

Bydel ID tilhører figur 5.1: Alna (1), Bjerke (2), Frogner (3), Gamle Oslo (4), Grorud (5), Grünerløkka (6), Nordre Aker (7) Nordstrand (8), Sagene (9), St.Hanshaugen (10), Stovner (11), Søndre Nordstrand (12), Ullern (13), Vestre Aker (14), Østensjø (15).

Disse grafene viser at det er vesentlige forskjeller i kvadratmeterprisen innad i Oslo-by. Grafene viser tydelig mønstre, som varierer utfra hvilken bydel man ser på. Figur 5.1 viser bydelenes spredning i priser for tidsrommet og et gjennomsnittspunkt som trekker frem de gjennomsnittlige forskjellene mellom bydelenes kvadratmeterpris. Figur 5.2 viser alle datapunktene til kvadratmeterprisen i de 15 bydelene, sammen med den gjennomsnittlige kvadratmeterprisen for Oslo som også vises i figur 5.4. Figur 5.3 viser utviklingen til styringsrenten og boliglånsrenten. Denne utviklingen ser vi på i sammenheng med prishoppet som skjedde i 2016. Her ser vi både fra figur 5.2, 5.4 og 5.5 hvordan den lave renten i 2016 bidro til enorm prisvekst. I figur 5.5 har vi laget et linjeplot for kvadratmeterprisen i alle bydeler. Her ser vi bydel 3 (Frogner) som i snitt har de høyeste kvadratmeterprisene, og bydel 12 (Søndre Nordstrand) med de laveste kvadratmeterprisene.



Figur 5. 5 Gjennomsnittlig kvadratmeterprisutvikling for hver bydel

I disse figurene kan vi tydelig se prishoppet som skjedde i 2016. Grunnet oljeprisfall ble renten da justert ned, og styringsrenten lå derfor på 0,5% over en lang periode for å stimulere økonomien. Det var da et begrenset antall boliger på markedet, og dette førte til kraftig prisvekst. Videre ser vi fra 2017 at prisene roet seg ned, noe som kan ha vært en effekt av boliglånsforskriften som ble innført 01.01.2017. Denne forskriften gjorde det

vanskeligere å anskaffe en sekundærbolig, da egenkapitalkravet for sekundærboliger i Oslo ble oppjustert til 40%.

5.1.2 Paneldata-resultater

Før vi presenterer resultatene fra paneldataanalysen vil vi beskrive testresultater og presisere valg vi tar for endelig modell.

5.1.2.1 Test for heteroskedastisitet, autokorrelasjon og tverrsnittsavhengighet

Vi har foretatt tester i Stata for heteroskedastisitet, autokorrelasjon og tverrsnittsavhengighet da dette må korrigeres for om det er til stede. Først foretok vi test for autokorrelasjon, denne heter «Wooldridge test for autocorrelation in panel data». Nullhypotesen i testen er fravær av autokorrelasjon. Med en p-verdi lik 0, forkaster vi null. Testen for heteroskedastisitet heter «Modified Wald test for groupwise heteroskedasticity». Nullhypotesen i testen er homoskedastisitet. P-verdien her var også lik 0. Den siste testen vi har kjørt er test for tverrsnittsavhengighet. Dersom tverrsnittsavhengighet er til stede i modellen og vi ikke justerer for dette vil det føre til villedende resultater (Baltagi, H. B., 2003). Her har vi kjørt to tester som forteller oss at residualene i modellen er korrelerte på tvers av tverrsnittene. Disse testene heter «Breusch-Pagan Lagrange Multiplier test of independence» og «Pesaran's test of cross-sectional independence». Testene indikerer at tverrsnittsavhengighet er til stede. I slike tilfeller foreslår Hoechle, D. (2007) at man bruker Driscoll and Kraay standard feil i modellen. Denne error-strukturen er antatt å være heteroskedastisk, autokorrelert opp til en viss *lag*, og sannsynligvis korrelert mellom panelene. Driscoll-Kraay standardfeil er robuste for veldig generelle former for tverrsnitt og tidsmessig avhengighet når tidsdimensjon, t er større enn, i. Siden testen for autokorrelasjon viser at vi har første ordens autokorrelasjon, legger vi inn “en” *lag* i Driscoll-Kraay modellen.

5.1.2.2 Robust Hausman-test

Basert på testresultatene ovenfor har vi gjennomført en robust Hausman-test. Ettersom paneldataen er heteroskedastisk og autokorrelert kan vi ikke bruke en vanlig Hausman-test

(Kaiser, 2014). I denne testen kan vi ikke forkaste nullhypotesen, og testen indikerer derfor at vi bør bruke en RE-modell.

5.1.2.3 Valg av modeller

I henhold til studiets formål mener vi at der foreligger faste bydelsforskjeller. Derfor faller valget på FE, selv om den robuste Hausman-testen indikerer at vi bør bruke RE. En FE modell vil i forhold til vår problemstilling kunne gi mer informasjon og forklare mer for førstegangskjøpere. Det foreligger heller ikke store forskjeller i estimatene til modellene. Siden en RE-modell gir estimater på hvordan tilfeldige effekter påvirker kvadratmeterprisen i de forskjellige bydelene, ser vi det som mer hensiktsmessig å estimere faste effekter.

Videre vurderer vi om modellen skal ha logaritmisk form for snitt av kvadratmeterpris, snitt av p-rom og snitt av soverom. Ved å bruke log av disse variablene kan vi beholde nullhypotesen om homoskedastisitet i *«Modified Wald test for groupwise heteroskedasticity»*. Problemet som da oppstår er at AR(12) variabelen faller bort, da flere av datapunktene har minusverdi. Her ønsker vi at modellen skal få med seg sesongvariasjoner, og kunne forklare kvadratmeterprisen ut ifra forrige års kvadratmeterpris. Vi velger derfor å ikke gjøre om variablene til logaritmisk form, slik at vi kan beholde AR(12).

I forhold til testresultatene i paneldataen, gjør vi modellvalg i forhold til robusthet. Som beskrevet innledningsvis i oppgaven ønsker vi en enkel modell. For å korrigere for heteroskedastisitet, autokorrelasjon og tverrsnittsavhengighet, finnes det flere modellvalg som kan produsere robuste standardfeil. Vi går her for en enkel løsning og bruker Driscoll-Kraay standardfeil. Denne modellen krever et panel med mange *t* verdier hvor formålet er å beregne standardfeil som er robuste for romlig korrelasjon og seriell korrelasjon.

5.1.2.4 Fixed effects-modell med Driscoll-Kraay standardfeil

Alle paneldata-modellene vi tester har samme avhengige variabel, men noen av modellene har forskjellige forklaringsvariabler. I FE-modellene tester vi først med bare boliglånsrente og AR12 som forklaringsvariabel, for så å legge til de resterende variablene for å se hvordan den faste effekten endrer seg. Boliglånsrenten og AR12 varierer kun over tid, så resultatene for disse rendyrker virkningen av resterende variablene i forhold til å fange opp variasjon på

tvers av bydelene. Vi tester også uten robuste standardfeil, for sammenligning. Alle modellene vi tester har tilfredsstillende F-verdier på 0,0, som tilsier at alle koeffisientene i modellene ikke er lik null.

Snitt av kvmpris	(1) Fixed Effects	(2) Fixed Effects Driscoll-Kraay	(3) Fixed Effects	(4) Fixed Effects Driscoll-Kraay
Boliglånsrente	-14605.3***	-14605.3***	-14418.3***	-14418.3***
	(190.3)	(578.6)	(187.2)	(557.1)
AR12	0.205***	0.205***	0.149***	0.149*
	(0.0252)	(0.0572)	(0.0251)	(0.0652)
Snitt av antall soverom			9806.3***	9806.3***
			(1556.1)	(1612.6)
Snitt av p-rom			-919.7***	-919.7***
			(82.31)	(110.8)
_cons	104294.3***	104294.3***	138220.8***	138220.8***
	(625.5)	(1986.0)	(3158.5)	(4826.5)
N	1800	1800	1800	1800
R^2	0.777	0.777	0.793	0.793

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Tabell 5. 2 Paneldataresultater for hovedmodell

FE-modell 1 og 2 analyserer her den gjennomsnittlige kvadratmeterprisens faste effekter med boliglånsrente som uavhengig variabel.

Estimatene fra denne modellen er «within» transformasjonen, og viser variablenes verdier hvor snittet fra hver bydel over tid er fratrukket for den avhengige og de uavhengige variablene. I modell 1 og 2 fra tabell 5.2 vises koeffisienten til boliglånsrenten og AR12 med og uten robuste standardfeil. Koeffisientene er signifikant i begge med en p-verdi på 0.00, og standardfeilen øker betraktelig i den standardrobuste modellen. Koeffisienten til boliglånsrenten forteller oss at 1% økning for boliglånsrenten har en negativ effekt på kvadratmeterprisen. Koeffisienten til AR12 forteller oss at gjennomsnittlig kvadratmeterpris

for ett år siden forklarer nåværende pris med 20,5%. Konstantleddet er også signifikant med p-verdi på 0.00. Resultatene fra modell 1 og 2

I modell 3 og 4 fra tabell 5.2 vises modellen med alle forklaringsvariabler med og uten robuste standardfeil. Her er den faste effekten for boliglånsrenten relativt lik som i modell 1 og 2, men effekten avtar noe. Videre ser vi at snitt av antall soverom har en positiv innvirkning på prisen, som forteller oss den faste effekten av ett soverom i kvadratmeterprisen. Snitt av p-rom viser seg å ha en negativ innvirkning på kvadratmeterprisen. Modellen forteller oss at større boliger har lavere kvadratmeterprisen. Videre ser vi at AR12 har en positiv innvirkning på kvadratmeterprisen på 14.9%. Dette forteller oss at kvadratmeterprisen på tidspunkt $P_{i,t-12}$ har en fast positiv effekt på kvadratmeterprisen i tidspunkt $P_{i,t}$. Forskjellene mellom modell 3 og 4 ligger i standardfeilen. Modell 3 er ikke justert for diagnosene som framkom i testene, så resultatene her kan være forventningsforskjøvet. Modell 4 legger her til høyere standardfeil, som resulterer i mindre t-verdier. I modell 3 og 4 er alle variablene signifikante. Alle variablene har en p-verdi på 0.00 der AR12 har en p-verdi på 0.024 i modell 4. Dette kommer av *lag*-strukturen som legges til i modellen. Denne lag-strukturen setter vi til én *lag*, siden testen for autokorrelasjon viste at vi har første ordens autokorrelasjon.

Modellenes R^2 er høy og forteller oss hvor mye av variasjonen til den gjennomsnittlige kvadratmeterprisen som forklares av de uavhengige variablene. Her ser vi at boliglånsrenten og AR12 alene fanger opp størsteparten av variasjonen.

5.1.2.5 Alternativ paneldata-modell

I hovedmodellen ser vi på den faste effekten over ti år. Her vet vi at boligmarkedet i Oslo har endret seg mye over tidsperioden og at ulike forhold i bydelene endrer seg over tid.

Vi velger derfor å lage en alternativ modell med et kortere tidsintervall. Her velger vi å analysere tidsperioden 2016-2019. Denne tidsperioden er kortere og nærmere dagens situasjon, og kan derfor gi et mer korrekt bilde av de faste effektene. Her utelater vi koronaåret 2020 på grunn av unormale tilstander i førstegangskjøpermarkedet. Den alternative paneldata-modellen bruker også Driscoll-Kraay standardfeil.

Snitt av kvmpris	(1) Fixed Effects	(2) Fixed Effects Driscoll-Kraay	(3) Fixed Effects	(4) Fixed Effects Driscoll-Kraay
Boliglånsrente	9719.4***	9719.4***	9115.1***	9115.1***
	(915.4)	(2579.7)	(848.7)	(2217.0)
AR12	0.268***	0.268***	0.226***	0.226***
	(0.0191)	(0.0477)	(0.0181)	(0.0523)
Snitt av antall soverom			8941.8***	8941.8***
			(1622.4)	(2484.4)
Snitt av p-rom			-929.6***	-929.6***
			(88.41)	(110.9)
_cons	44130.8***	44130.8***	81433.2***	81433.2***
	(2326.8)	(6736.0)	(4013.6)	(6347.0)
N	720	720	720	720
R^2	0.299	0.299	0.403	0.403

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Tabell 5. 3 Paneldataresultater for alternativ modell

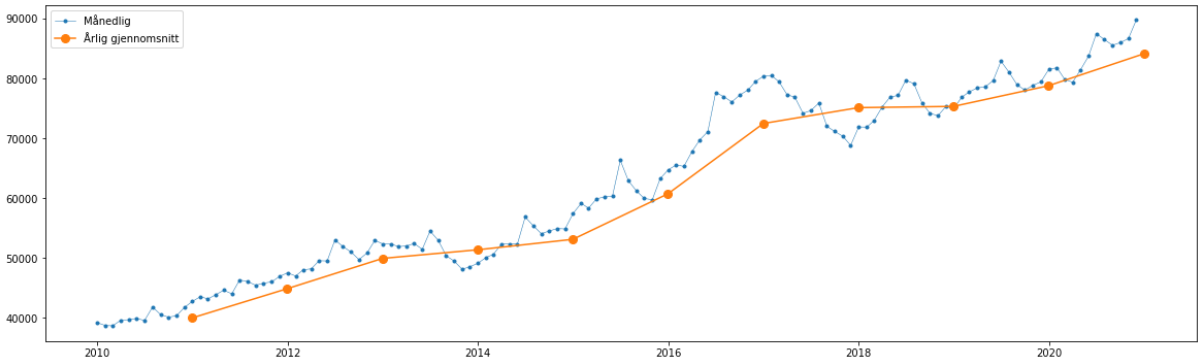
5.1.2.6 Paneldataresultater for alternativ modell

I den alternative modellen som vises i tabell 5.3 er tidshorisonten forkortet til 2016-2019. Resultatene her viser at den faste effekten til variablene fortsatt har samme fortegn som i hovedmodellen, men at koeffisientene har endret seg noe. Vi ser at den faste effekten til boliglånsrenten har sunket og AR12 har steget. Boliglånsrenten og AR12 forklarer langt mindre av variasjonen til kvadratmeterprisen enn hva den gjør i hovedmodellen. Snitt av antall soverom gir fortsatt en positiv effekt, men har avtatt noe. Snitt av p-rom har fortsatt negativ effekt, og AR12 har økt til 22.6%. Alle forklaringsvariablene og konstantleddet er her signifikant med p-verdier på 0.00.

5.2 Resultater fra tidsserieanalysen

Vi presenterer i denne delen hvordan vi har gått frem for å bygge vår SARIMA modell og legger frem dens resultater. Modellen er bygget for å kunne predikere den gjennomsnittlige kvadratmeterprisen i Oslo 1,5 år frem i tid. Vi foretar først to prediksjoner av data vi allerede

har, for å se hvor godt modellen treffer. Vi benytter månedlige kvadratmeterpriser fra 1.1.2010 til 31.12.2020, som tilsvarer 132 tidsperioder. Den avhengige variabelen i vårt datasett noteres med *y* og er den gjennomsnittlige kvadratmeterprisen for Oslo.



Figur 5. 6 Kvadratmeterprisens trend og sesongvariasjoner

Når vi studerer datagrunnlaget i figur 5.6, kan vi tydelig se at det er en trend i datasettet. Vi ser både en trend over tid og en sesongvariasjon. Denne trenden og variasjonene fjerner vi ved å differensiere.

Trenden fjerner vi ved å foreta en første ordens differensiering. Dette gjør vi ved å trekke *y*-1 fra *y*. Første-differensiering blir da:

$$y_{t-1} = y - y_{t-1} \tag{5.1}$$

Vi foretar så en Augmented Dickey-Fuller test (ADF-test) for å se om datasettet er stasjonært. Testens nullhypotese er at dataen ikke er stasjonær. I vårt tilfelle får vi en p-verdi på 0,021 og vi forkaster null. Selv om datasettet er stasjonært innenfor et 95% konfidensintervall, ser vi at tidsserien fortsatt har tendenser til sesongvariasjon (se vedlegg).

For å få datasettet ytterligere stasjonært innenfor et 99% konfidensintervall foretar vi også en sesong differensiering. Denne kan utledes slik:

$$y_{t-12} = y - y_{t-12} \tag{5.2}$$

Ny ADF-test gir oss en p-verdi på 0,003 og vi forkaster nullhypotesen, se vedlegg.

5.2.1 Valg av parameter basert på ACF, PACF-plots og grid search

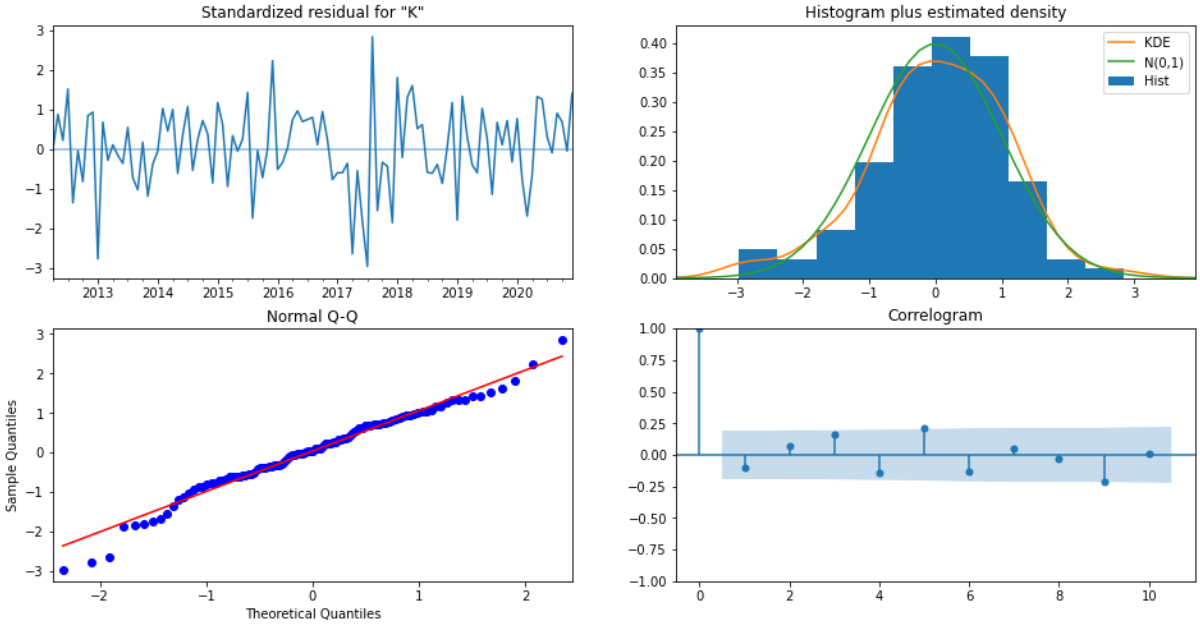
SARIMA modeller er bygget opp av hyperparameterne (p, d, q)*(P, D, Q)m. For å finne hvilke verdier parameterne skal ha, benytter vi ACF- og PACF-plottet. Vi benytter den differensierte tidsserien og fra ACF-plottet ser vi at en *lag* struktur på 1 eller 2 er en mulighet, se vedlegg. Det samme gjelder for PACF-plottet, også her er det en *lag* struktur på 1 eller 2, se vedlegg.

Av dette følger at alle parameterne for p, P, q og Q kan være mellom 1 eller 2, vi vet også at d og D begge er 1 siden dette ga stasjonæritet.

Vi benytter oss av en liste med de mulige parameterne og får grid-search modellen til å finne den kombinasjonen som gir lavest AIC-score. Modellen gir oss følgene parameterverdier: SARIMA (1, 1, 1)*(0, 1, 1, 12).

Modell diagnoser

For at resultatene i denne analysen skal være valide må tidsserieregresjons antakelser være møtt. Her ser det ut til at vi møter alle antakelsen, med resultater i figur 5.7.



Figur 5. 7 Diagnose plott for SARIMA-modellen

Øverst i venstre hjørne ser vi residualene over tid. Disse ser ut til å være uten noen sesongvariasjon eller trend og vi kan dermed anta at modellen har klart å fange opp dette på en tilfredsstillende måte. Plottet viser at residualene i snitt ligger rundt null, med en konstant varians. Øverst til høyre ser vi hvordan linjen for Kernel Density Estimation (KDE) legger seg i forhold til linjen for normalfordeling rundt 0 og standardfeil 1 (N(0,1)). Det er ønskelig at disse linjene ligger tett på hverandre da dette indikerer at residualene er normalfordelte. Nede til venstre ser man fordelingen av residualene (blå prikker) følge en lineær trend (rød linje) av utvalgene tatt fra en standard normalfordeling med snitt på 0 og standardavvik på 1. Til sist ser man nederst til høyre ett «correlogram». Dette viser at residualene har liten korrelasjon med de laggete versjonene av seg selv, presentert ved at de blå prikkene ligger innenfor det skraverte området.

Basert på disse resultatene kan vi konkludere at residualene er nær normalfordelt med konstant varians, og at vi har en modell som er godt tilpasset vårt datasett.

SARIMA-modeller

SARIMA-modellen vi bygger benytter seg av de parameterne vi fant i forrige avsnitt. Vi deler datasettet opp slik at perioden fra 1.1.2010 til 1.5.2018 fungerer som treningsdata og perioden fra 1.6.2018 til 1.12.2020 fungerer som testdata. Dette betyr at SARIMA-modellen benytter seg av treningsdatasettet til å estimere modellen, for så å lage prediksjoner fra 1.6.2018. SARIMA-modellen er spesifisert ved SARIMA(p, d, q)*(P, D, Q, s), der vår modell spesifiseres SARIMA (1, 1, 1)*(0, 1, 1, 12) og kan uttrykkes på formen:

$$P_t = a\,P_{t-1} + b\,\varepsilon_{t-1} + d\,\varepsilon_{t-12} + \varepsilon_t \qquad (5.3)$$

Der a = ar.L1, b = ma.L1 og d = ma.S.L12

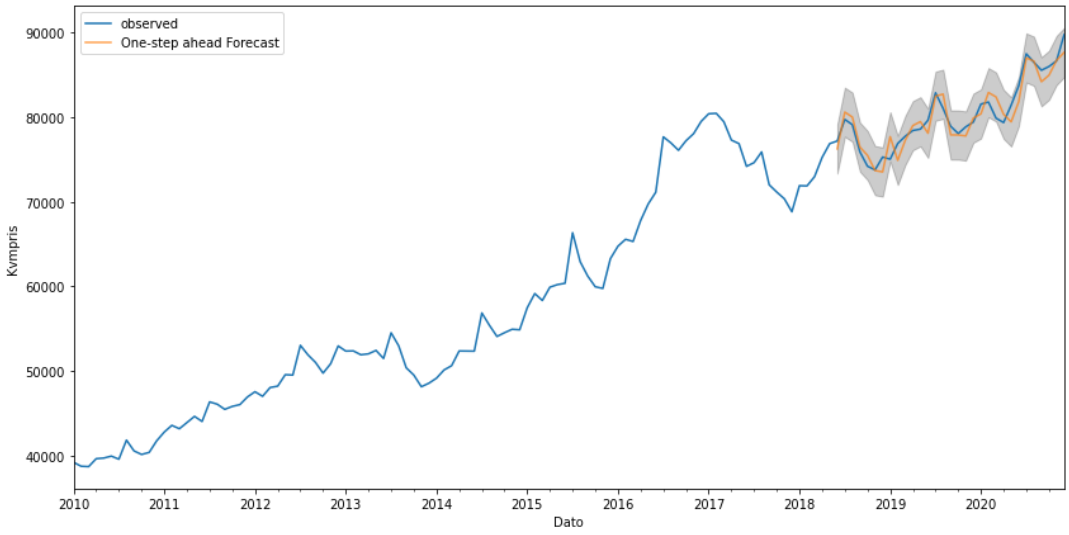
Resultatet av denne modellen ser man i tabell 5.4.

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.7364	0.194	3.794	0.000	0.356	1.117
ma.L1	-1.7555	0.744	-2.358	0.018	-3.215	-0.296
ma.S.L12	-0.4509	0.046	-9.887	0.000	-0.540	-0.362
sigma2	7.14e+05	6.03e+05	1.184	0.236	-4.68e+05	1.9e+06

Tabell 5. 4 SARIMA-resultater

Tabellen viser at de valgte parameterne i modellen er signifikante med p-verdier under 0,05, og vi kan dermed anta at vi har funnet en modell som passer godt.

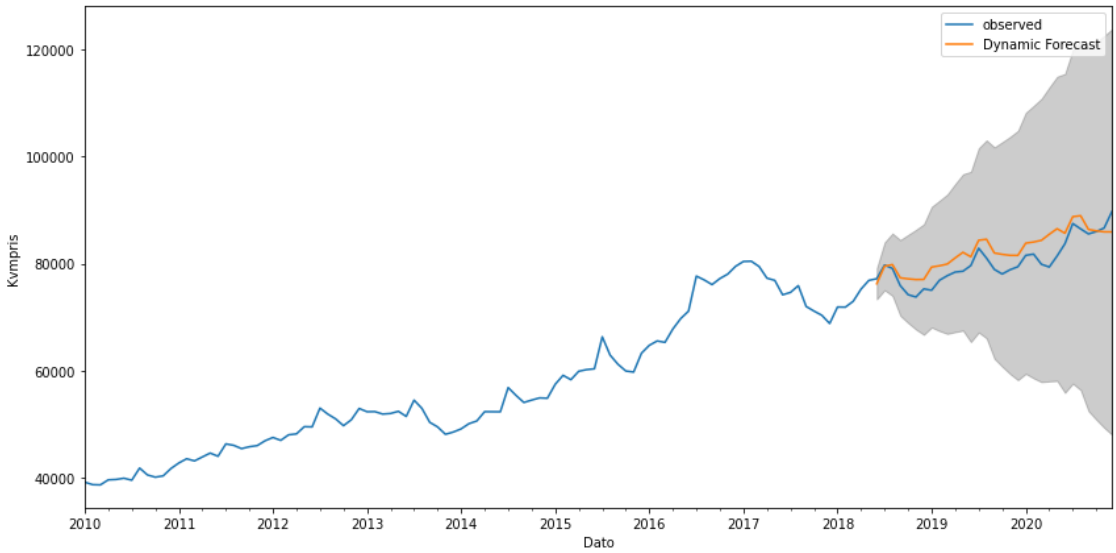
Vi benytter oss så av to ulike fremgangsmåter når vi tester modellen. Først lager vi en «One-step ahead» prediksjon. Her bruker modellen alle datapunkter fra 2010m1 frem til 2018m5 til å predikere neste måned. Måneden etter blir så predikert ved hjelp av alle de tilgjengelige datapunktene opp til tidspunktet for prediksjonen. Modellen fortsetter slik helt til hele testdatasettet er predikert. Denne fremgangsmåten gir ofte gode resultater siden modellen hele tiden innhenter ny informasjon. Dette ser vi i figur 5.8 nedenfor.



Figur 5. 8 One-step ahead

I grafen er datasettet i sin helhet representert med en blå linje og den oransje linjen representerer de predikerte verdiene. Vi har også lagt til et 95% konfidensintervall representert med et skravert felt. Prediksjonen har en RMSE score på 1301,46.

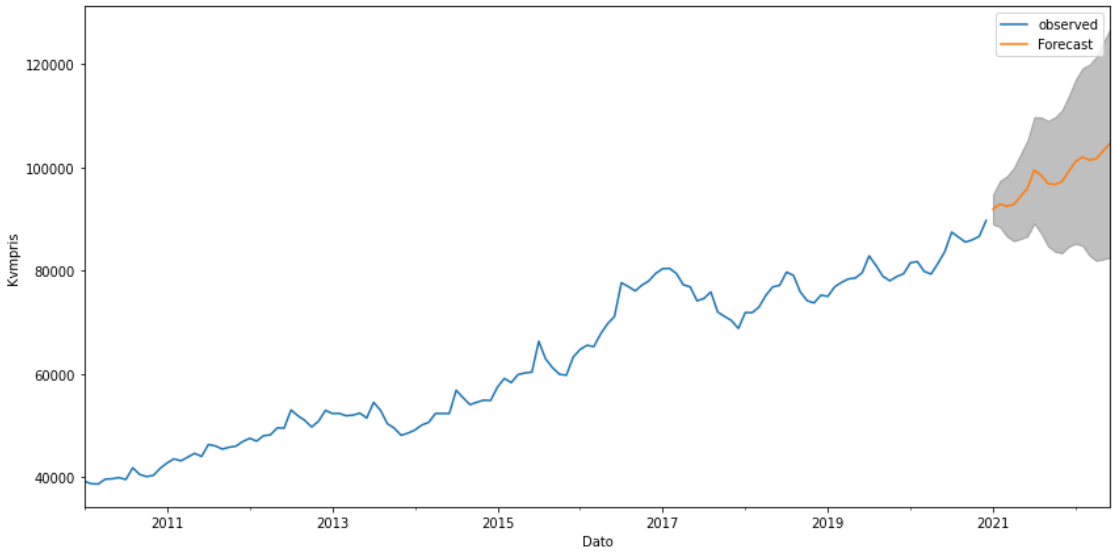
Videre benytter vi oss av en dynamisk modell. På samme måte som med «one-step-ahead» predikerer vi 1.6.2018 basert på treningsdatasettet. Forskjellen mellom de to modellene er at den dynamiske predikerer frem til 31.12.2020 uten å innhente ny informasjon. Dette ser vi i figur 5.9.



Figur 5. 9 Dynamisk prediksjon

Den dynamiske modellen får en RMSE score på 2833,83, som er vesentlig høyere enn for «one-step-ahead». Vi ser at også denne klarer å fange opp sesongvariasjonen og den stigende trenden.

For å predikere boligprisene 1,5 år frem i tid benytter vi oss av den innebygde prognosefunksjonen i Statsmodels biblioteket. Denne funksjonen gjør at vi kan bruke modellen vi har laget til å predikere fremtiden. Vi benytter oss av den samme fremgangsmåten som tidligere, men vi utvider datasettet til å inkludere alle tilgjengelige verdier. Prognosen kan derfor sammenlignes med den dynamiske modellen siden vi ikke her får innhentet ny informasjon med faktiske verdier som i «one-step-ahead». I figur 5.10 vises SARIMA-modellens prognose for kvadratmeterpris i Oslo frem til 1.6.2022.



Figur 5. 10 Dynamisk prognose fra 1.1.21-1.6.22

Resultatene til prognosemodellen viser en relativt høy stigning av kvadratmeterprisen frem mot juni 2021. Resultatene her vil diskuteres nærmere i neste kapittel. I vedlegg ligger prognosens predikerte verdier, samt verdiene for øvre og nedre grense i konfidensintervallet.

5.3 Resultater fra prediksjonsanalysen med regresjonsalgoritmer

Resultatene fra maskinlæringsanalysen blir målt i henhold til måltallene vi presenterer i kapittel 4. Resultatene blir presentert samlet i en tabell, før vi presenterer den beste modellens «Feature Importance».

5.3.1 Maskinlæringsresultater

Resultatene til modellene vises i tabell 5.3 Benchmark modellen vår er et enkelt valgtre. Vi har valgt dette fordi den er den enkleste modellen av maskinlæringsmodellene vi benytter. Videre ser vi resultatene til Random Forrest og Extreme Gradient Boosting modellen.

Modell	MAE	MSE	R2	RMSE
Desicion trees	5390,40	54741623,83	0.873	7398,76
Random Forrest	3037,12	18600856,68	0.957	4312,87
XGBoost	2602,87	12680904,72	0.971	3561,03

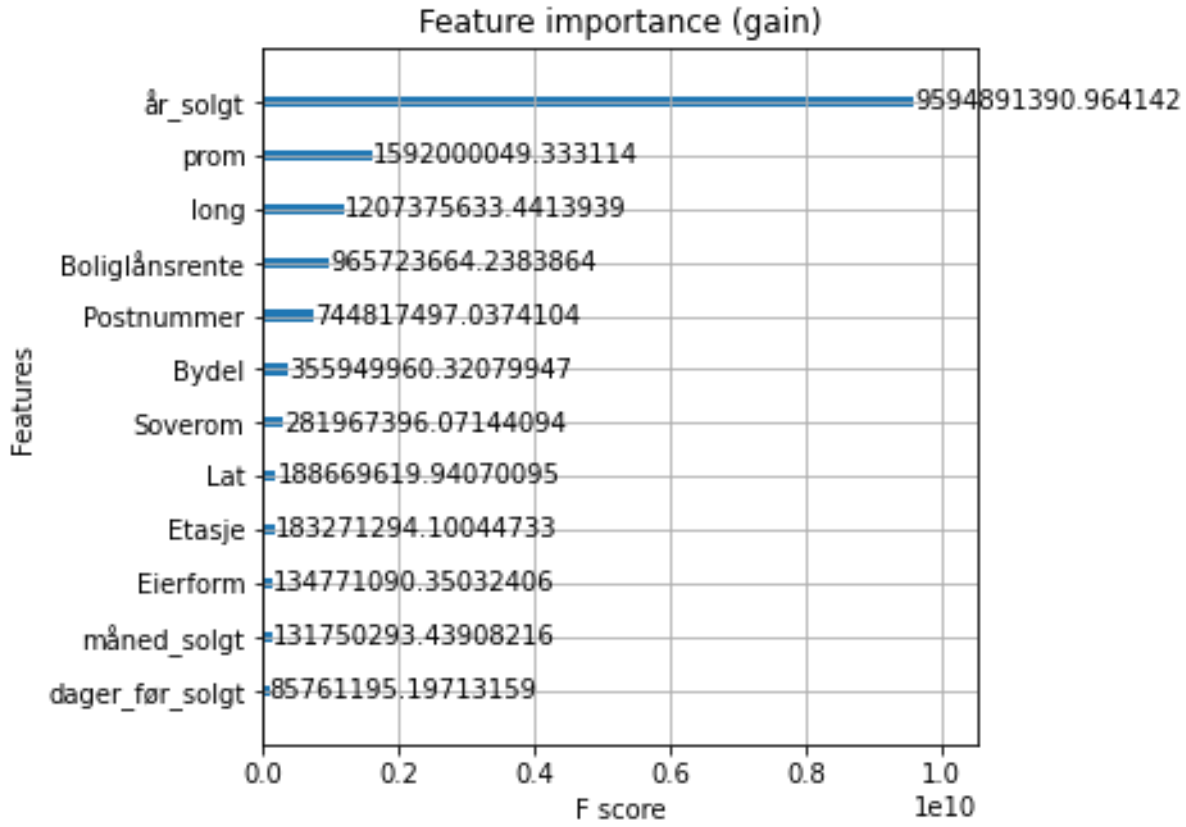
Tabell 5. 5 Maskinlæringsresultater

Fra tabell 5.5 ser vi at XGBoost-modellen predikerer best, med en RMSE på 3561,03. Dette tilsier at den beste modellen vår predikerer et gjennomsnittlig standardavvik mellom faktiske og predikerte verdier på 3561kr per kvadratmeter. XGBoost har også lavest MSE som tilsier at modellen har lavest varians mellom predikerte og faktiske verdier. MAE er også lavest for XGBoost og forteller oss at snittet til residualene er på 2602,87 kr per kvadratmeter.

5.3.2 Feature importance

Sett i sammenheng med oppgavens problemstilling, presenterer vi her XGBoost-modellens Feature importance.

Dette er en metode som vektlegger variablene, ut ifra hvor gode de er til å predikere kvadratmeterprisen i XGBoost-modellen. Vi får dermed innsikt i hvilke variabler denne modellen utpeker som mest og minst viktige for prediksjon av kvadratmeterpris. Innenfor Feature importance funksjonen i XGBoost-modeller er det flere aspekter som gjør at modellen vekter variabler som viktige eller uviktige. Da denne modellen er kompleks velger vi for enkelhets skyld å bruke den feature importance attributten som heter «Gain». Vi bruker denne fordi den er mest relevant til å måle den relative nytten til hver enkelt variabel, eller «feature» som det kalles innen maskinlæring. «Gain» beskriver det relative bidraget en variabel gir modellen ved å regne hver variabels bidrag for hvert beslutningstre i XGBoost. En høyere verdi av denne beregningen sammenlignet med en annen variabel betyr at modellen vekter den som en bedre prediktor.



Figur 5. 11 Feature Importance

I figur 5.11 ser vi hvordan XGBoost vekter de forskjellige variablene. Vi kan tydelig se at hvilket år leilighet er solgt har stor betydning for modellen. Vi diskuterer resultatene herfra nærmere i neste kapittel.

6. Diskusjon

Diskusjonskapittelet deles inn i tre deler hvor vi diskuterer resultater og funn i de tre analyse, med svakheter og anbefalinger for videre forskning.

6.1. Diskusjon av paneldata resultater

Resultatene fra hovedmodellen med faste effekter viser at variasjonen i den gjennomsnittlige kvadratmeterprisen i stor grad kan forklares med endringer i boliglånsrenten. Modellen forteller at en renteøkning på 1% trekker kvadratmeterprisen ned. Denne nedgangen kan forklares ved at en renteheving begrenser konsumet til befolkning som har eller skal ta opp boliglån. En renteheving begrenser derfor pengebruken i markedet noe som kan påvirker tilbud og etterspørsel. Dette resultatet forteller oss også at en rentenedgang øker konsumet, noe som har motsatt effekt. I koronaåret 2020 har vi sett effekten av rentenedgang. Med en styringsrente på 0%, og lave boliglånsrenter har boligmarkedet i Oslo vokst kraftig.

Videre viser resultatene at boliger med et adskilt soverom hever kvadratmeterprisen betydelig. Boligene vi analyserer er små og har for det meste to soverom. Snittet her ligger på rett over ett soverom og mange av boligene ingen adskilte soverom. Resultatet forteller oss at verdien av et soverom kontra en bolig uten adskilt soverom har stor betydning for kvadratmeterprisen til boligene. Dette gjelder også for boliger med to soverom. Med to soverom har eventuelle førstegangskjøpere muligheten til å leie ut et rom, og dermed lettere betjene boliglån og andre kostnader. Videre ser vi at snitt av p-rom har en negativ påvirkning på kvadratmeterprisen. Dette betyr at jo mindre boligen er jo mer betaler man per kvadratmeter. Totalt sett koster en 35kvm leilighet mindre enn en 50kvm leilighet, så dette ser ut til å være et resultat av den store etterspørselen av mindre og «billige» leiligheter. Siden førstegangskjøpere har begrenset tilgang på kapital vil de kjøpe den boligen de har råd til, og derfor blir de nødt til å betale betydelige prispåslag per kvadratmeter. Videre har vi AR12 variabelen. Denne gir en indikasjon av sesongvariasjonen i kvadratmeterprisen, og forteller oss at kvadratmeterprisen delvis kan forklares av den gjennomsnittlige kvadratmeterprisen for ett år siden.

Hovedmodellen tar for seg tidsperioden mellom 2011 til 2020. I denne tidsperioden har det skjedd mye i Oslo, hvor bydeler har endret seg og økonomien har endret seg. Resultatene i hovedmodellen må derfor tas med en klype salt, siden bydelene og leilighetsmarkedet ikke er faste over tid. Vi forsøker med den alternative modellen å se på en kortere tidsperiode som er nærmere dagens situasjon. Dette gjør vi for å se om de faste effektene har endret seg. Resultatene fra den alternative paneldatamodellen viser at R^2 er mye lavere enn den er i hovedmodellen. Dette kan komme av at med en kortere tidsperiode, får vi ikke med like mye variasjon i den gjennomsnittlige kvadratmeterprisen. Der boliglånsrenten forklarer så og si alt av variasjonen i hovedmodellen har boliglånsrenten mindre forklaringskraft i den alternative modellen. Det skal nevnes her at boliglånsrente ble satt ned i 2016 grunnet oljeprisfall, for å stimulere økonomien. Dette førte til høy boligprisvekst, som etter hvert avtok i 2017. Selv om prisene avtok, ble renten holdt lavt på samme nivå i denne tidsperioden. Det kan her tenkes at boliglånsforskrifter som ble innført i 2017, hadde stor effekt. Det kan derfor diskuteres at den alternative modellen ikke får like høy R^2 fordi det lave boliglånsrentenivået ikke forklarer prisendringene som skjedde i prisnedgangsperioden i 2017.

I henhold til studiens problemstilling kan fixed-effects modellen gi førstegangskjøpere et innblikk i hvordan variablene har faste effekter på kvadratmeterprisen over tid. Modellen kan gi førstegangskjøpere et visuelt og teoretisk bilde av hvordan renten er med på å bestemme prisutviklingen for boligmarkedet. Paneldata-modellene er også med på å beskrive hvordan antall kvadratmeter og soverom er med på å påvirke kvadratmeterprisen. Her er førstegangskjøperboligene relativt unike ved at små leiligheter driver kvadratmeterprisen opp. Sånn sett forteller modellen at man betaler mindre per kvadratmeter dersom leiligheten er større. Førstegangskjøpere er som tidligere beskrevet en prissensitiv kjøpergruppe, der mange kjemper om de mindre leilighetene. Med mindre leiligheter blir da totalprisen lavere og derfor mer populær for kjøpere med mindre kjøpekraft. Når det kommer til AR12 variabelen, gir denne en indikasjon på at det finnes en fast effekt som kan forklare at kvadratmeterpris er avhengig av prisen for et år siden.

Når det kommer til prisforskjeller mellom bydelene og hva som forårsaker dette, har vi ikke kunnet dra noen konklusjoner. Resultatene forteller oss at noe ved disse bydelene fører til at

den gjennomsnittlige kvadratmeterprisen varierer. Modellen forteller ikke hva som forårsaker variasjon, men at det er uobserverte individspesifikke effekter som fører til at den gjennomsnittlige kvadratmeterprisen varierer.

Paneldata-modellene gir et relativt intuitivt bilde av hvordan boligmarkedet fungerer, men den har svakheter som må adresseres.

6.1.1 Svakheter ved paneldataanalysen og forslag til videre forskning

Svakhet i modellens robusthet – vi har i denne masteroppgaven valgt å ikke fokusere på optimalisering av modell. Vi vet at modellen kunne vært mer robust dersom vi la mer fokus på modellvalg. Vi har hatt som mål å vise til at det finnes signifikante faste effekter som beskriver kvadratmeterprisutviklingen. Vi foreslår at videre forskning undersøker andre paneldatamodeller, da dette kan gi et mer robust resultat. Vi vet fra den robuste Hausman-testen at random-effects muligens hadde vært en bedre modell. Vi anbefaler derfor å undersøke om en tilfeldig effekt modell, kan være et bedre utgangspunkt for analyse.

Bydelspanelene er sammensatt av flere gjennomsnittlige verdier, og disse har liten variasjon som vist i sammendragsstatistikken. Snitt av antall soverom og snitt av kvadratmeter har gjennomsnittsverdier som varierer lite, og modellen kan ha derfor svakheter her.

Da deler av denne studien har basert seg på bydelsanalyser av Oslo, har vi kunnet trekke frem variasjon og forskjeller i hver bydel. En svakhet med dette er at Oslo-boligmarkedet kan variere mye i pris innad hver bydel. Vi får derfor ikke fanget opp nabolagsvariasjoner. Vi vet at i de mest attraktive bydelene kan kvadratmeterprisen variere mye, og det er her vi ikke får fanget sentrale og spesifikke effekter som er med på å drive prisene. Vi anbefaler derfor å undersøke om en nabolagsanalyse er mulig å gjennomføre.

Videre anbefaler vi å undersøke om det finnes flere variabler som kan forklare den faste effekten til kvadratmeterpris. Her kan for eksempel en kontrollvariabel for endringer i boliglånsforskrifter prøves ut. Det hadde også vært interessant å se om kollektivtilbud, kulturelle fasiliteter og tilgang på naturområder rundet byen har en effekt.

6.2 Diskusjon av tidsserieanalyse

Resultatene fra tidsserieanalysen deles her inn i tre deler, hvor vi først diskuterer den fremtidige prognosen. Deretter diskuterer vi resultatene til den dynamiske modellen og deretter one step ahead-modellen.

Resultatene til den fremtidige prognosemodellen viser at kvadratmeterprisen kommer til å stige mye. Her kan det oppstå problemer ved å velge et fast tidsrom for prognose. Hendelser i denne tidsperioden kan ha stor påvirkning på treffsikkerheten til modellen. Prognoser i tidsperioder med store prissvingninger kan derfor lede til andre resultater enn i perioder med stabilitet. Resultater og testing har også vist at starttidspunkt for fremtidig prognose har stor betydning for modellens presisjon. Dersom starttidspunktet er i en periode med en voksende trend, har modellen en svakhet i at den følger denne trenden. Vår fremtidige prognose frem til 01.06.2021 er derfor preget av denne svakheten. Kvadratmeterprisene til oppgavens avgrensning har i løpet av 2020 vokst kraftig i perioder, og modellen vår ser ut til å følge denne utviklingen.

Resultatene til den dynamiske modellen viser at prediksjonene treffer relativt bra. Modellen har en fordel ved at den kan predikere flere perioder frem i tid og gi en indikasjon på hvordan prisene vil bevege seg. En svakhet ved denne modellen er at treffsikkerheten er lavere. Som vist i resultatkapittelet har modellen et skravert område som viser til et konfidensintervall på 95%. Modellen treffer derfor prisen innenfor dette område med 95% sannsynlighet. Her er det altså sannsynlighet for at modellen kan falle utenfor, og bomme kraftig. Som vist i resultatkapittelet vil en lengre periode øke konfidensgrensen. Så den dynamiske modellen vil kunne treffe langt dårligere i slutten av perioden, enn hva den gjør i starten. Uansett treffer vår modell relativt bra på det testdatasettet vi gir den.

Til sammenligning har One step ahead modellen en mindre konfidensgrense. Konfidensintervallet er også her 95%, men siden modellen kun predikerer et steg frem i tid øker ikke konfidensgrensen. One step ahead har derfor en klar fordel i at den har høyere sannsynlighet for å treffe prisen. Dette viser også resultatet, ved at prediksjonen følger den faktiske prisen med en RMSE på 1301. En ulempe med denne modellen er at kun predikerer et steg frem.

Kjøp av leilighet innebærer beslutningen «kjøpe nå» eller vente en tidsperiode. En vesentlig ulempe med å «kjøpe nå» er at man reduserer framtidig handlingsrom etter som det er kostnader (dokumentavgift, betaling til megler mm.) man ikke får tilbake ved videresalg. Slik sett er prognoser en periode fram i tid kanskje det perspektivet som passer best med beslutningssituasjonen til de fleste som vurderer å kjøpe bolig.

For førstegangskjøpere har SARIMA-modellen flere fordeler. Ved å analysere sesongvariasjonene og prissvingningene kan en førstegangskjøper se hvilke måneder som er populære og upopulære. Da prisene i boligmarkedet er preget av tilbud og etterspørsel kan en slik analyse være nyttig. Dersom en vil kjøpe i en periode der markedet har lav etterspørsel kan man i beste fall legge inn et lavt bud, og unngå budkrig som kan driver prisen opp. Resultatene forteller også at førstegangskjøpere kan dra nytte av å være fleksible i når de velger å innta markedet. Selv om disse slutningene kan tas fra våre modeller vil vi poengtere at modellene kan ta feil. Å følge med på månedlig prisstatistikk samt historiske prissvingninger kan gi like mye, hvis ikke mer innsyn. Her kan modellene dog brukes til å predikere neste periodes pris med relativt høy nøyaktighet, men langtidsprognoser vil være risikabelt.

6.2.1 Svakheter i tidsserieanalysen og anbefalinger for videre studier

I henhold til svakheter i våre modeller vil vi her komme med anbefalinger for videre forskning. For den dynamiske modellen vil vi anbefale å teste om et «rullende vindu» kan gi bedre presisjon. Med rullende vindu, bruker modellen et fast «vindu» den baserer prediksjonen på. I stedet for å bruke hele datahistorikken, settes vinduet til for eksempel 2010m1 til 2018m6. Prosessen beveger seg så videre ett steg (en måned), og predikerer til slutt 2020m12 med «in sample» perioder fra 2011m12 til 2020m11. Med rullende vindu fjernes observasjonene fra starten av datasettet når prosessen flytter seg steg for steg. Siden sesongvariasjoner og observasjoner fra 2010 ikke nødvendigvis har stor betydning for hvordan prisen varierer i 2019 og 2020, kan et rullende vindu gi bedre resultater. Dette gjelder spesielt i situasjoner der det har skjedd strukturelle endringer over tid, f.eks. som følge av eksterne sjokk til økonomien. Det før omtalte oljeprisfallet er et eksempel på et slikt eksternt sjokk.

Å velge riktig lengde for prognosen og vinduer er vilkårlig. Her er det et kompromiss mellom å velge for få perioder og å bruke en lang periode, som kan gi enda mer tilfeldige resultater. Vi anbefaler her at videre forskning undersøker rullende vindu med ulike lengder på vinduet.

Siden tidsserieanalysen med SARIMA baserer seg på månedlig gjennomsnittlig kvadratmeterpris, har dette medført at mye av informasjonen i datasettet er vasket ut. For å unngå dette anbefaler vi å gjennomføre en lignende analyse for hver enkelt bydel. Man vil da kunne se hvordan prisene har variert innad i byen, og få en bedre oversikt over kvadratmeterpriser på bydelsnivå.

Videre anbefaler vi å teste forskjellige parameterverdier i modellen. Der vi bruker AIC med grid-search, kan man også bruke BIC eller analysere lag strukturen på egenhånd. En samlingsanalyse av disse metodene kan være interessant for å se hvilken fremgangsmåte som gir høyest presisjon.

En siste anbefaling er å bruke andre prognosemodeller for å sammenligne treffsikkerheten. Vi anbefaler å teste en automatisk univariat prognosemodell som heter Prophet. Denne modellen er utviklet av Facebook og er tilgjengelig i Python. Den skal fungere best på tidsseriedata med mye sesongvariasjon og flere sesonger med historisk data (Taylor. S.J., Letham. B., 2017). Videre anbefaler vi å teste modeller som tar med andre forklarende variabler. Vi ser at tidsserien alene fort kan bomme på prediksjoner i perioder med unormale prissvingninger. Vi tror derfor en modell med makroøkonomiske faktorer eller andre variabler som har en påvirkning på kvadratmeterprisen kan treffe bedre i slike perioder. Her vil vi anbefale modeller som Vektor Autoregresjon (VAR), eller Autoregressiv distribuert lag (ADL). VAR er en prognosemodell som brukes når to eller flere tidsserier har påvirkning på hverandre. Modellen krever to eller flere tidsserievariabler, der disse variablene også bør ha en påvirkning på hverandre. Grunnen til at modellen heter autoregresjon, kommer av at hver variabel er modellert som en funksjon av foregående verdier. VAR-modellen er en fortsettelse av AR-modellen med flere variabler. I VAR-modellen har variablene både en påvirkning på hverandre og er påvirket, så det er ikke noen foretrukken y variabel. ADL-modellen kombinerer en AR(p) modell med laggede verdier av forklarende variabler.

6.3 Diskusjon av maskinlæringsanalysen

Målet i maskinlæringsanalysen har vært å produsere en modell som kan predikere kvadratmeterpris med høy presisjon. Utover dette har vi hatt som mål å se hvorvidt vi kan dra slutninger i hvorfor modellen predikerer som den gjør, og se hvilken nytte en slik modell har for førstegangskjøpere. Resultatene fra maskinlæringsanalysen viser at både Random Forest og Extreme Gradient Boosting predikerer kvadratmeterprisen med høy presisjon. Når det kommer til innsyn i XGBoost-modellen er det vanskelig å trekke slutninger. Som først antatt er det vanskelig å si hvorfor modellen predikerer som den gjør, og se hvilke variabler som beskriver kvadratmeterprisen best.

På bakgrunn av vår kompetanse og erfaring innen prediksjon med maskinlæring, kunne modellen vært bedre. Da dette er et nytt fagfelt for oss, så vi en mulighet til å tilnærme oss ny kunnskap. Vi vet derfor at det finnes forbedringspotensialer i denne analysen.

Modellen vår er bygd ved at den predikerer 20% av datasettet vårt, og bruker 80% til å lære seg dynamikken og mønstre i datasettet. Denne inndelingen er tilfeldig, så modellen predikerer kvadratmeterprisen til et tilfeldig utvalg av boligtransaksjoner som strekker fra 2010-2020. Denne metoden er enkel og en av grunnene til at vi velger den. En annen grunn er at 80/20 inndeling gir modellen mulighet til å lære av historisk data. På den måten lærer modellen seg pristrenden i tidsperioden, og kan derfor skille kvadratmeterpris for like boliger i forskjellige tidsperioder. Her kan modellen ha en svakhet i at den legger like mye vekt på å predikere kvadratmeterpriser i 2010 som den gjør i 2020. Modellen har derfor ikke fokus på å predikere kvadratmeterprisen i et nærliggende tidsrom. Så måltallene til modellen forteller oss hvor mye den i snitt bommer fra faktiske priser i hele tidsperioden. Vi kan derfor ikke med sikkerhet si at modellene vil ha like resultater dersom man kun predikerer kvadratmeterpris for boliger i 2020.

Det skal også nevnes at lignende modeller ofte brukes til å predikere totalprisen til en bolig og ikke kvadratmeterprisen. For at oppgaven skulle ha en gjennomgående variabel å analysere, falt valget på kvadratmeterpris. Her ser vi i ettertid at predikert kvadratmeterpris gir mindre bruksverdi enn hva totalpris hadde gjort.

For førstegangskjøpere kan XGBoost-modellen brukes til å estimere kvadratmeterpris, gitt at man har tilgang til modellens variabler. Dette kan være nyttig dersom man vil sammenligne den estimerte kvadratmeterprisen med gjennomsnittlige kvadratmeterpriser for lignende og nærliggende boliger. Her kan det tenkes at estimert totalpris hadde vært mer nyttig og mer intuitivt. Vi må også nevne at denne modellen ville være nyttig dersom man kontinuerlig oppdaterer datasettet, slik at modellen kan lære med nye priser og boligdata.

Når det kommer til innsyn og forklaring i hvorfor modellen predikerer som den gjør, har vi som antatt hatt problemer med å dra slutninger. Vi har her valgt metoden som gir oss «Gain» i feature importance. Denne metoden forteller hvilke variabler modellen vekter som viktigst for å predikere med nøyaktighet. Det kan se ut til at hvilket år boligen er solgt har mest å si for prediksjonen. Videre er antall kvadratmeter, lengdegrad, boliglånsrente, postnummer, bydel, soverom, breddegrad, etasje, eierform og måned solgt, representert som viktige variabler. XGBoost-modellen er komplisert, og vi kan ikke med sikkerhet si at vår tolkning av modellen er korrekt. Det finnes flere metoder å evaluere variablene i modellen, men å forstå alle disse metodene faller utenfor vår forståelse og kompetanse. Vi trekker derfor vi ingen konklusjoner herfra, men forstår hvorfor de fremhevede variablene er viktige for prediksjonen.

Siden boligprisene i datasettet har en klar trend, er det forståelig at modellen vekter årstall høyest. Videre ser vi at antall kvadratmeter vektes høyt, som vi vet har stor forklaringskraft på kvadratmeterpris. Modellen vekter også geografiske variabler høyt, som vi med sikkerhet kan si er utslagsgivende variabler for kvadratmeterpris i Oslo. Her er lengdegraden rangert høyest som forteller om boligen ligger i øst eller vest. Videre rangeres boliglånsrenten før boligens postnummer, bydel og breddegrad. Som vist i paneldataanalysen er boliglånsrenten med på å forklare variasjon i kvadratmeterpriser, så vi forstår hvorfor modellen vekter variabelen høyt. Til slutt ser vi soverom, etasje, eierform og måned solgt. For soveroms variabelen vet vi fra paneldataanalysen at antall soverom har stor påvirkning på kvadratmeterprisen. Kjellerleiligheter og boliger på bakkeplan prises ofte lavere enn høyere etasjer og det kan tenkes at modellen har funnet et mønster som priser etasjer forskjellig. Når det kommer til eierform vet vi at selveier ofte er dyrere da man blant annet har mer frihet for utleie. Boligene vi analyserer er typiske investeringsobjekter for utleie, og det kan

tenkes at dette fanges opp av modellen. Til slutt har vi måned solgt, som kun har en tallverdi fra 1-12 for hvilken måned boligen ble solgt. Dette gir en indikasjon om at modellen fanger opp sesongvariasjonen.

6.3.1 Svakheter i maskinlæringsanalysen og anbefalinger for videre forskning

Som nevnt ovenfor er inndelingen av treningsdata og testdata gjort på en enkel 80:20 inndeling, der utvalget er tilfeldig. Vi anbefaler å teste ut andre måter å inndele treningsdata og testdata. For at modellen skal ha en verdi for brukere er det viktig at den predikerer med høy nøyaktighet i datasettets siste periode. Vi anbefaler derfor å teste en lignende modell på datasettets siste perioder, for å se om resultatene endres.

Vi har i denne oppgaven valgt å teste ut tre populære maskinlæringsalgoritmer innen regresjon. Vi har ikke sikkert grunnlag til å anta at dette er de beste. Vi anbefaler derfor å utforske om andre modeller kan oppnå høyere treffsikkerhet. Fra lignende studier i litteraturgjennomgangen har vi sett at modeller som nevrale nettverk, Super Learner og Elastic Net, har oppnådd gode resultater. Elastic Net kombinerer to lineære modeller og er en regresjonsmodell med regularisering. Ensemble modellen Super Learner er en samlemodell og Neural Network som er kunstige nevrale nettverk. De nevnte modellene representerer populære og typiske tilnærminger for sine grupper. Innen hver av disse gruppene finnes det flere varianter av hver modell. Vi anbefaler derfor å teste disse modellene for å se om de kan gi høyere treffsikkerhet.

Verdsetting av boliger med maskinlæring er typisk forklart med en hedonisk prismodell. Datagrunnlaget i dette studiet har flere karakteristikker som beskriver boligene, og passer derfor den hedoniske teorien, utenom boliglånsrenten. Selv om datagrunnlaget dekker viktige egenskaper til boligene, er det flere karakteristikker som kunne styrket våre modellers treffsikkerhet. Vi anbefaler derfor å innhente flere datavariabler om det er mulig. Her lister vi variabler som kunne gitt modellen høyere treffsikkerhet:
solforhold, oppussingsgrad, renovasjon, tilstandsgrad, utsikt, veranda/balkong, heis, garasje/parkering, antall bad, antall toaletter, lenge og breddegrad, tilgang på kulturelle fasiliteter og tilgang på kollektivtransport.

Utenom datavariabler som beskriver boligkarakteristikker, hadde det vært interessant å undersøke hvorvidt buddata har en påvirkning på prisen. Antall budgivere og antall interessenter per bolig, kan være et interessant utgangspunkt for videre forskning.

Her vil vi også anbefale en annen del innenfor maskinlæring. En klassifiseringsmodell kan være interessant for å predikere om salgsprisen er lik, over eller under prisantydning. Her kan man analysere om det finnes trekk og kjennetegn ved boliger eller buddata som fører til at boliger selges over eller under prisantydningen.

7. Konklusjon

I denne studien har vi undersøkt hvordan paneldata med faste effekter, univariat tidsserieanalyse med SARIMA og prediksjon med regresjonsalgoritmer innen maskinlæring kan brukes til å analysere kvadratmeterpriser. Med utgangspunkt for førstegangskjøpere i Oslo, forsøker vi å trekke frem fordeler og ulemper ved analysemetodene. Vi gjentar studiens problemstilling: *Hvilke fordeler og ulemper har metodene paneldata, tidsserie og prediksjon med regresjonsalgoritmer, for analyse av boligpriser?*

I paneldataanalysen kan vi konkludere med at det finnes faste effekter som påvirker kvadratmeterpris på bydelsnivå. Fordelen med denne analysen er at den fremhever faktorer som påvirker kvadratmeterprisen over tid. Her ser vi at endringer i boliglånsrenten er høyt korrelert med endringer i boligprisen. En ulempe med paneldataanalyse er begrensninger i antall bydelsvariabler. Det er med andre ord flere faste effekter modellen tillegger feilleddet, som vi ikke vet hva er. Modellen har også en ulempe i at bydeler endrer seg over tid, noe som må tas i betraktning for de estimerte faste effektene. Videre har modellen en fordel i at den på en oversiktlig måte kan vise prisnivået i de 15 bydelene over tid, men en ulempe i at vi ikke kan trekke ut faste effekter fra hver bydel.

Den univariate tidsserieanalysen med SARIMA har fordeler i at den enkelt kan predikere kvadratmeterpris med relativt høy treffsikkerhet. Modellen har vist seg å være preget av starttidspunkt for prediksjon, og kan derfor være mindre treffsikker i perioder med unormale prissvingninger. Der den dynamiske SARIMA modellen har en fordel i å kunne predikere flere perioder frem i tid, ligger det en risiko i at sannsynligheten for å bomme øker jo lenger prediksjonen er. Her har "One step ahead" modellen en fordel i at den er mer treffsikker, men ulempe i at den kun predikerer en periode frem i tid. Utover treffsikkerhet i prediksjoner gir analysen innsyn i kvadratmeterprisens sesongvariasjon.

I prediksjonsanalysen bruker vi tre maskinlæringsalgortimer som heter Decision Trees, Random Forest og Extreme Gradient Boosting. Av disse tre predikerer XGBoost modellen best og oppnår høy treffsikkerhet. Modellen har en klar fordel i at den kan predikere kvadratmeterpris med høy treffsikkerhet basert på det historiske datasettet med tilhørende boligattributter og variabler. Ulempen i denne analysen er innsyn i hvorfor

maskinlæringsmodellen predikerer som den gjør. Her forsøker vi å trekke frem variabler som modellen vekter som viktige, men grunnet modellens kompleksitet drar vi ingen konklusjoner herifra.

Samlet gir analysene et overordnet blikk for hvordan boligpriser endrer seg, og hvorfor prisene endrer seg. Vi kan trekke frem endringer i boliglånsrenten som en nøkkelvariabel i å forklare prissvingninger. Vi ser også at boligenes beliggenhet er en viktig variabel til å forklare prisforskjeller.

Analysene våre illustrerer sterke og svake sider ved de ulike metodiske tilnærmingene. Kjøp av bolig er en av de viktigste økonomiske beslutningene folk flest foretar seg. De økonomiske konsekvensene for den enkelte boligkjøper kan være store. Dette tilsier at personer som vurderer å kjøpe bolig kan ha økonomisk gevinst ved å innhente ekspertråd på hvilke faktorer som påvirker boligprisene. Analyser av boligmarkedet er kompliserte, og vi opplever at vi har lært mye, men enda ikke nok om boligmarkedet og egnede analysemetoder.

For de som vurderer å kjøpe bolig er det viktig å være klar over at boligkjøp er litt som å «cashe» inn en opsjon, dvs. det er elementer av irreversibilitet i beslutningen om å kjøpe bolig: Har man først kjøpt med de ekstra ikke refunderbare kjøpskostnadene som dokumentavgift, står man svakere økonomisk rustet viss en interessant bolig skulle dukke opp på markedet litt senere. Dette gjelder kanskje spesielt for førstegangskjøpere som ofte har høy gjeldsgrad og få ubundne økonomiske ressurser for å gjøre om et klart ikke-optimalt boligkjøp.

8. Referanseliste

Baltagi, B. H. (2003). *Econometric Analysis of Panel Data.* 3. Utg. Chichester: John Wiley & Sons Ltd. Tilgjengelig fra: https://himayatullah.weebly.com/uploads/5/3/4/0/53400977/baltagi-econometric-analysis-of-panel-data_himmy.pdf (Lest: 09.06.2021)

Benedictow, A. *et al.* (2020a) *Vanskeligstilte på boligmarkedet og betydningen av et velfungerende langsiktig leiemarked. Rapport 20-2020.* Samfunnsøkonomisk analyse AS. Tilgjengelig fra: *https://static1.squarespace.com/static/576280dd6b8f5b9b197512ef/t/5f0d600dea043118c9729179/1594712081637/R20-2020+Vanskeligstilte+p%C3%A5+boligmarkedet+og+betydningen+av+et+velfungerende+langsiktig+leiemarked.pdf (*Lest: 18 .05.2021)

Benedictow, A. et al. (2020b) *Skatt i den norske boligmodellen*. Rapport 26-2020. Samfunnsøkonomisk analyse AS. Tilgjengelig fra: https://static1.squarespace.com/static/576280dd6b8f5b9b197512ef/t/5f623dab5784a345eff38197/1600273855278/R26-2020+Skatt+i+den+norske+boligmodellen.pdf (Lest: 09.03.2021)

Boliglånsforskriften. *Boliglånsforskriften 1. januar 2020–31. desember 2020.* Tilgjengelig fra: https://www.regjeringen.no/no/tema/okonomi-og-budsjett/finansmarkedene/boliglansforskriften-1.-januar-202031.-desember-2020/id2679449/

Bolstad, E. (2009-2021). *Postnummer i Oslo kommune*. Tilgjengelig fra: https://www.erikbolstad.no/postnummer-koordinatar/kommune.php?kommunenummer=301#forklaring (Lest 27.03.2021)

Breiman, L. (1996) *Bagging Predictors.* Tilgjengelig fra: https://link.springer.com/content/pdf/10.1023/A:1018054314350.pdf (Lest 04.04.2021)

Driscoll, J, C., & Kraay, A, C,. (1998) Consistent Covariance Matrix Estimation with Spatially Dependent Panel Data, *Review of Economics and Statistics,* 80 (4): 549-560. Tilgjengelig fra: https://EconPapers.repec.org/RePEc:tpr:restat:v:80:y:1998:i:4:p:549-560 (Lest: 27.04.2021)

Chen, X. & Wei, L. & Xu, J. (2017). *House Price Prediction Using LSTM.* Tilgjengelig fra: https://arxiv.org/ftp/arxiv/papers/1709/1709.08432.pdf, (Lest: 15.03.2021)

Eiendom Norge (u.å.,a*) Den norske boligmodellen*. Tilgjengelig fra: *https://eiendomnorge.no/om-oss/visjon-og-verdier/den-norske-boligmodellen* Lest: 26.04.2021)

Eiendom Norge (u.å.,b*) Om statistikken*. Tilgjengelig fra: *https://eiendomnorge.no/boligprisstatistikk/om-statistikken/ (*Lest: 07.04.2021)

Eiendomsmegler.no (2019) *Styr unna obligasjonsleiligheter*. Tilgjengelig fra: https://eiendomsmegler.no/obligasjonsleilighet (Lest 22.05.2021)

Eiendomsverdi. (2019) *Den norske sykepleierindeksen 2019.* Tilgjengelig fra: https://eiendomnorge.no/blogg/den-norske-sykepleierindeksen-2019-article360-923.html (Lest 07.07.21)

Feng, Y. & Jones, K. (2015). *Comparing Multilevel Modelling and Artificial Neural Networks in House Price Prediction*. Tilgjengelig fra: https://www.researchgate.net/publication/275967043_Comparing_Multilevel_Modelling_and_Artificial_Neural_Networks_in_House_Price_Prediction (Lest: 15.03.2021)

Finanstilsynet (2020) *Boliglånsundersøkelsen 2020. Dok. Nr. 16/2020.* Tilgjengelig fra: https://www.finanstilsynet.no/nyhetsarkiv/pressemeldinger/2020/boliglansundersokelsen-2020/ (Lest 19.05.2021)

Forskning.no (2019), *Derfor er nordmenn verdensmestere i å pusse opp.* Tilgjengelig fra: https://forskning.no/hus-og-hjem-okonomi/derfor-er-nordmenn-verdensmestere-i-a-pusse-opp/1573237 (Lest: 03.03.2021)

Fortmann-Roe, S. (2012a). *Accurately Measuring Model Prediction Error*. Tilgjengelig fra: http://scott.fortmann-roe.com/docs/MeasuringError.html (Lest 7.4.2021)

Fortmann-Roe, S. (2012b). *Understanding the Bias-Variance Tradeoff.* Tilgjengelig fra: http://scott.fortmann-roe.com/docs/BiasVariance.html (Lest 7.4.2021)

Friedman, J. (2001) *Greedy function approximation: A gradient boosting machine.* Tilgjengelig fra: https://projecteuclid.org/journals/annals-of-statistics/volume-29/issue-5/Greedy-function-approximation-A-gradient-boostingmachine/10.1214/aos/1013203451.full (Lest: 06.06.2021)

Glorot, X. & Bengio, Y. (2010) *Understanding the difficulty of training deep feedforward neural networks.* Tilgjengelig fra: https://proceedings.mlr.press/v9/glorot10a.html (Lest: 26.03.2021)

Hoechle, D., (2007) Robust Standard Errors for Panel Regressions with Cross-Sectional Dependence, *The Stata Journal*, 7(3), pp. 281–312. doi: 10.1177/1536867X0700700301.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2017). *An introduction to statistical learning, volume 112.* NewYork: Springer: Tilgjengelig fra: https://link.springer.com/content/pdf/10.1007%2F978-1-4614-7138-7.pdf (Lest.4.2021)

Jorly, J. (2020) *Xgboost — in a nutshell.* Tilgjengelig fra: https://ai.plainenglish.io/xgboost-in-a-nutshell-211e170e8b48 (Lest: 06.06.2021)

Glen, S. (2019) *Decision Tree vs Random Forest vs Gradient Boosting Machines: Explained Simply*. Tilgjengelig fra: https://www.datasciencecentral.com/profiles/blogs/decision-tree-vs-random-forest-vs-boosted-trees-explained (Lest 04.04.2021)

Hastie, T., Tibshirani, R. & Friedman, J. (2008), *The Elements of Statistical Learning.* Tilgjengelig fra: https://web.stanford.edu/~hastie/Papers/ESLII.pdf (Lest 29.03.2021)

Holmes, S. (2000). *RMS Error. Stanford University.* Stanford, USA. Tilgjengelig fra: http://statweb.stanford.edu/~susan/courses/s60/split/node60.html (Lest 07.07.2021)

Jacobsen, D. H., & Naug, B. E. (2004) Hva driver boligprisene?. *Penger og kreditt,* 4: 229-240. Tilgjengelig fra: https://www.norges-bank.no/globalassets/upload/publikasjoner/penger_og_kreditt/2004-04/jacobsen.pdf (lest 10.07.21)

James, G., Witten, D. Hastie, T, Tibshirani R. (2013) *An Introduction to Statistical Learning: With Applications in R*. Springer. 1. Utg. New York: Springer. Tilgjengelig fra: https://doi.org/10.1007/978-1-4614-7138-7

Kaiser, B., (2014). RHAUSMAN: Stata module to perform a (cluster-)robust Hausman test, University of Bern.

Krogsveen (2021). *Jeg vil kjøpe leilighet. Hvilken eierform passer best for meg?* Tilgjengelig fra: https://www.krogsveen.no/magasin/leilighet-ulike-eieformer-hva-betyr-det-for-deg-som-kjoper (Lest 23.07.2021)

Kuhn, M., Johnson, K. (2013). *Applied Predictive Modeling*. Tilgjengelig fra: https://link.springer.com/content/pdf/10.1007%2F978-1-4614-6849-3.pdf (Lest 29.03.2021)

Lee, C. (2021) *Predicting land prices and measuring uncertainty by combining supervised and unsupervised learning*. Tilgjengelig fra: https://journals.vgtu.lt/index.php/IJSPM/article/view/14293 (Lest 30.03.2021)

Li, S. et al., (2021). Understanding the Effects of Influential Factors on Housing Prices by Combining Extreme Gradient Boosting and a Hedonic Price Model (XGBoost-HPM). Land, 10 (5): 533. doi: http://dx.doi.org/10.3390/land10050533 (lest 15.08.21)

Milunovich, G. (2019*). Forecasting Australian Real House Price Index*: A Comparison Study of Machine Learning and Time Series Methods. Tilgjengelig fra: https://www.researchgate.net/publication/334388950_Forecasting_Australian_Real_House_Price_Index_A_Comparison_Study_of_Machine_Learning_and_Time_Series_Methods (Lest 29.03.2021)

Mullainathan, S., & Spiess, J. (2017). *Machine Learning: An Applied Econometric Approach.* Tilgjengelig fra: https://www.aeaweb.org/articles?id=10.1257/jep.31.2.87 (Lest: 06.04.2021)

Nielsen, A. (2019). *Practical Time Series Analysis: Prediction with Statistics and Machine Learning. O'Reilly Media, Inc.* Tilgjengelig fra: https://shop.aer.io/oreilly/p/practical-time-series/9781492041658-9149

Mullainathan, S., Spiess, J. (2017). *Machine Learning: An Applied Econometric Approach. Journal of Economic Perspectives,* 31 (2): 87-106. doi: 10.1257/jep.31.2.87 (lest 15.04.21)

Norske Boligbyggelags Landsforbund SA (NBBL) (2020), *Dramatisk nedgang i førstegangskjøpernes kjøpekraft*. Tilgjengelig fra: https://www.nbbl.no/aktuelt/12-02-2020-dramatisk-nedgang-i-forstegangskjopernes-kjopekraft/ (Lest: 06.05.2021)

NRK.no (2021) *Færre unge kjøper bolig: Tram (22) får ikke boliglån i Bodø – må ha over en mill i egenkapital.* Tilgjengelig fra: https://www.nrk.no/nordland/xl/faerre-unge-kjoper-bolig_-tram-_22_-far-ikke-nok-boliglan-i-bodo-_-ma-ha-over-en-mill.-i-egenkapital-1.15387275 (Lest: 13.06.2021)

Oslo kommune (u.å.), *Boligpriser*. Tilgjengelig fra: https://www.oslo.kommune.no/statistikk/boliger-byggevirksomhet-arbeids-og-naringsliv/boligpriser/ (Lest 15.03.2021)

Oslomet (2020a). *Dette bestemmer boligprisene.* Tilgjengelig fra: https://www.oslomet.no/forskning/forskningsnyheter/dette-bestemmer-boligprisene (lest 5.08.21)

OsloMet (2020b). *Norwegian Housing Market Watch 2020*. Tilgjengelig fra: https://housinglab.oslomet.no/wp-content/uploads/2020/03/NorwegianHouseWatch_digital.pdf (Lest: 24.04.2021)

Pedace, R. (2013). *Econometrics for Dummies.* 1. Utg. New Jersey: John Wiley & Sons, Inc. Tilgjengelig fra: https://www.wiley.com/en-us/Econometrics+For+Dummies-p-9781118533871

Pedregosa et al. (2011) *Machine Learning in Python*, JMLR 12, pp. 2825-2830. Tilgjengelig fra: https://scikit-learn.org/stable/modules/tree.html#tree (Lest: 03.04.2021)

Rydne, N. og Alsberg, O. (2020) *Frykter ukontrollert prisvekst – ser 2016-tendenser i boligmarkedet.* Tilgjengelig fra: https://e24.no/naeringsliv/i/xPKEOV/frykter-ukontrollert-prisvekst-ser-2016-tendenser-i-boligmarkedet (Lest: 02.07.2021)

Røed Larsen, E,. & Sommervoll, D. E. (2004) *Hva bestemmer boligprisene?.* Samfunnsspeilet, 2: 10-17. Tilgjengelig fra: https://www.ssb.no/priser-og-prisindekser/artikler-og-publikasjoner/hva-bestemmer-boligprisene (lest 10.07.21)

Sodini, P. & Van Nieuwerburgh, S. & Vestman, R. & von Lilienfeld-Toal, U. (2021) *Identifying the Benefits from Home Ownership: A Swedish Experiment.* Swedish House of Finance Research Paper No. 16-11. Tilgjengelig fra: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2785741 (Lest: 06.07.2021)

Statistisk sentralbyrå (SSB.no) (2017) *Statistikk om boligpriser.* Tilgjengelig fra: https://www.ssb.no/priser-og-prisindekser/artikler-og-publikasjoner/statistikk-om-boligpriser (Lest: 06.04.2021)

Statistisk sentralbyrå (SSB.no) (2018) *Formuesulikheten øker*. Tilgjengelig fra: https://www.ssb.no/bygg-bolig-og-eiendom/faktaside/bolig (Lest: 24.05.2021)

Statistisk sentralbyrå (SSB.no) (2019a) *Hvorfor lager SSB prognoser for utviklingen i norsk økonomi?* Tilgjengelig fra: https://www.ssb.no/nasjonalregnskap-og-konjunkturer/artikler-og-publikasjoner/hvorfor-lager-ssb-prognoser-for-utviklingen-i-norsk-okonomi (Lest: 16. 03.2021)

Statistisk sentralbyrå (SSB.no) (2019b) *Vi bruker boligen som sparegris.* Tilgjengelig fra: *https://www.ssb.no/inntekt-og-forbruk/artikler-og-publikasjoner/vi-bruker-boligen-som-sparegris* (Lest 03.04.2021)

Statistisk sentralbyrå (SSB.no) (2019c) *Færre unge kjøper bolig.* Tilgjengelig fra: https://www.ssb.no/bygg-bolig-og-eiendom/artikler-og-publikasjoner/faerre-unge-kjoper-bolig (Lest: 09.04.2021)

Statistisk sentralbyrå (SSB.no) (2021a) *Renter i banker og kredittforetak*. Tilgjengelig fra: https://www.ssb.no/bank-og-finansmarked/finansinstitusjoner-og-andre-finansielle-foretak/statistikk/renter-i-banker-og-kredittforetak (Lest: 03.05.2021)

Statistisk sentralbyrå (SSB.no) (2021b) *Kvadratmeterpriser for eneboliger*. Tilgjengelig fra: https://www.ssb.no/priser-og-prisindekser/boligpriser-og-boligprisindekser/statistikk/kvadratmeterpriser-for-eneboliger (Lest: 21.05.2021)

Statistisk sentralbyrå (SSB.no) (u.å.) *Fakta om bolig.* Tilgjengelig fra: https://www.ssb.no/bygg-bolig-og-eiendom/faktaside/bolig (Lest: 22.04.2021)

Stewart, M. (2019) *The Actual Difference Between Statistics and Machine Learning.* Tilgjengelig fra: https://towardsdatascience.com/the-actual-difference-between-statistics-and-machine-learning-64b49f07ea3 (Lest: 02.06.2021)

Taylor, J. S., & Letham, B., (2017) *Forecasting at scale.* PeerJ Preprints*,* 5:e3190v2. doi: https://doi.org/10.7287/peerj.preprints.3190v2 (Lest 10.08.2021)

Theoblad, O. (2017) *Machine Learning For Absolute Beginners.* Tilgjengelig fra: https://bmansoori.ir/book/Machine%20Learning%20For%20Absolute%20Beginners.pdf (Lest: 03.04.2021)

USBL.no (u.å). *Borettslag og sameie, hva er forskjellen?* Tilgjengelig fra: https://www.usbl.no/beboer/forskjellen-pa-borettslag-og-sameier (Lest: 24.05.2021)

Wooldridge, J. M. (2019). *Introductory Econometrics: A Modern Approach*. 7. utg. Boston: Cengage.

Wang, C. & Wu, H. (2018). *A new machine learning approach to house price estimation. New Trends in Mathematical Science.* Tilgjengelig fra: https://www.researchgate.net/publication/330001343_A_new_machine_learning_approach_to_house_price_estimation . (Lest: 16.03.2021)

Wilson, I. & Paris, S. & Ware, J. & Jenkins, D. (2002) *Residential property price time series forecasting with neural networks*. Tilgjengelig fra: https://www.sciencedirect.com/science/article/pii/S0950705101001691 (Lest: 16.03.2021)

9. Vedlegg

9.1 Behandling av manglende verdier og data i primærdatasettet

I primærdatasettet finnes det transaksjoner med manglende verdier som vi her gjør rede for.

Postnummer:	• 10 transaksjoner med NULL-verdi er fjernet.
Pris:	• 1557 NULL variabler er fjernet. • Totalt 1562 variabler er fjernet hvorav 5 av disse hadde urovekkende lav pris og så ut til å være feil.
Prisantydning:	• 331 variabler med NULL-verdi. • 253 av de samme boligene har 0 dager på markedet og kan se ut til å være arv/kupp.
Eierform:	• En av boligtransaksjonene har ukjent eierform og fjernes. • 11 av boligene i datasettet er obligasjonsleiligheter. Disse fjernes fra datasettet da de utgjør en særdeles liten andel og er sjeldne.
Byggeår:	• 976 av boligene i datasettet har verdien 0 eller NULL for byggeår. Disse fjernes for å unnvike støy og redusert presisjon i modellene.
P-rom:	• 511 av boligene har NULL verdier for P-rom. • 36 av boligene har verdien 0 for P-rom. • 4 boliger har verdien 1 for P-rom.
BRA:	• 2376 av boligene har verdien NULL for BRA. • 200 av boligene har verdien 0 for BRA. • 2 av boligene har verdien 1 for BRA.

P-rom og BRA:	• 143 av boligene/ transaksjonene mangler datapunkter for både p-rom og BRA. Disse transaksjonene fjernes fra datasettet. • Der enten BRA eller p-rom mangler verdier, fjernes transaksjonen.
Etasje:	• 21479 av boligene har verdien NULL for etasje. Dette velger vi å tolke som etasjen 0. • To av transaksjonene har usannsynlig høy etasje. Vi velger å fjerne disse, da det mest sannsynlig er en feil.
Soverom:	• 15314 av boligene har verdien NULL for soverom. • Disse endres til verdien 0, da flere mindre leiligheter ikke har adskilte rom for soverom.
Tinglysningsdato:	• Denne variabelen fjernes fra datasettet, da denne variabelen ikke er med på å beskrive boligens verdi.

Tabell 9. 1 Justeringer og endringer i datasett

Bakgrunnen til at vi fjerner boligtransaksjoner der det mangler verdier for en eller flere variabler kommer av at vi ønsker et komplett datasett for alle analysene. Et komplett datasett er med på å styrke prediksjonene og analysene, da alle boligtransaksjonene har variabler med data som er mulig å analysere.

9.2 Interpolering og partiell-analyse av styringsrente og boliglånsrente

Datagrunnlaget vi bygger analysene på bruker den gjennomsnittlige renten for pant i bolig. Grunnet forskjeller i tidsrekkene til SSB, har vi kun fått tak i månedlige data for boliglånsrenten fra 2014-2020. Fra 2010-2013 har vi kun kvartalsvise data å basere oss på. For at hele tidsrekken skal ha det samme datagrunnlaget, har vi foretatt en partiell analyse

av den månedlige styringsrenten og kvartalsvise boliglånsrenten. Boliglånsrenten kan i de fleste omstendigheter estimeres og sammenlignes basert på styringsrenten.

Fra denne analysen har vi så estimert oss frem til en månedlig boliglånsrente basert på analysens kostandledd og koeffisient. Vi har så tatt gjennomsnittet av estimeringen og den kvartalsvise boliglånsrenten, for å få et estimat som er nærmere det faktiske datagrunnlaget. Med denne fremgangsmåten fikk vi gode estimater som får med mye av variasjonene, men som fortsatt har feilestimater på opptil 0,5 prosentpoeng i noen av månedene.

Vi har derfor også foretatt en interpolering av de kvartalsvis boliglånstallene. Med denne fremgangsmåten får vi med oss månedlige variasjoner som sammenlignet med partiell analyse har mindre feilmargin. Da interpoleringen følger den faktiske renten, og har mindre feil enn den partielle analysen har vi valgt å basere analysene på interpolering av kvartalsvis boliglånsrente fra 2010 til 2013.

9.3 Boliglånsforskrifter

Boliglånsforskriftene bestemmer hvor mye boligkjøper kan låne ut ifra betjeningsevne, gjeldsgrad og belåningsgrad. Disse forskriftene er satt i verk for å bidra til en mer bærekraftig utvikling av husholdningers gjeld, som i stor grad har betydning for norsk økonomi. Regjerningen har satt rammer for bankenes utlånspraksis, for å hindre at husholdninger har for mye gjeld.

Betjeningsevne går ut på at banken skal beregne kundens evne til å betjene lånet. Her skal bankene ta hensyn til kundens inntekt og relevante utgifter. I beregningen skal bankene legge til grunn en rente på 5% høyere enn det aktuelle rentenivået. Denne renteøkningen legges til grunn for å se om lånekunden har tilstrekkelige midler til å dekke normale utgifter til livsopphold, etter en renteøkning. Dette gjøres for å stressteste lånekunden og sikre at låntageren har en likviditetsbuffer. Dersom lånekunden ikke har tilstrekkelige midler, kan lånet kun innvilges innenfor bankenes fleksibilitetskvote.

Forskriften for gjeldsgrad setter en grense for hvor mye en kunde kan låne ut ifra kunden tjener. Her skal kundens samlede gjeld ikke overskride fem ganger årsinntekt. I denne beregningen skal den samlede gjelden være med og ikke bare lån med pant i bolig.

I beregning av betjeningsevne tar banken utgangspunkt i lånekundes årsinntekt, og trekker fra eventuell gjeld, som f.eks studielån. samlede gjeld delt på årsinntekt. Her kan den samlede gjelden ikke overskride fem ganger årsinntekt.

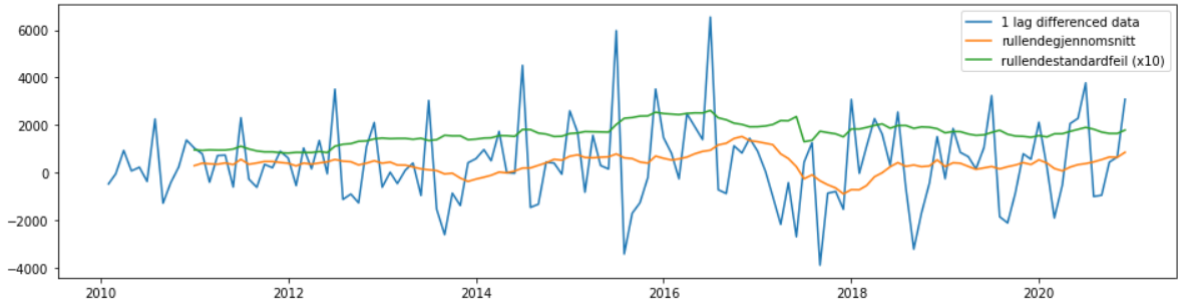
Belåningsgrad setter en grense på hvor mye en kunde kan låne i forhold til boligens verdi. Den maksimale belåningsgraden for nedbetalingslån er på 85%, så boligkjøper må ha egenkapital på minimalt 15%. For rammekreditter er den maksimale belåningsgraden satt til 60%. I beregninger av belåningsgrad skal alle lån med pant i boligen tas med, også fellesgjeld i borettslag og sameier. Boliglånsforskriftene endrer seg i tidsintervall på ca. 4-5 år og er der for å sikre at utviklingen i boligprisene er bærekraftige. I 2015 ble egenkapitalkravet for nedbetalingslån oppjuster fra 10% til 15%. Videre ble det i 2017 satt en 60% grense for maksimal belåningsgrad for sekundærboliger i Oslo (Regjeringen, 2021).

9.4 Tidsserieanalyse resultater og tester

Augmented Dickey-Fuller tester

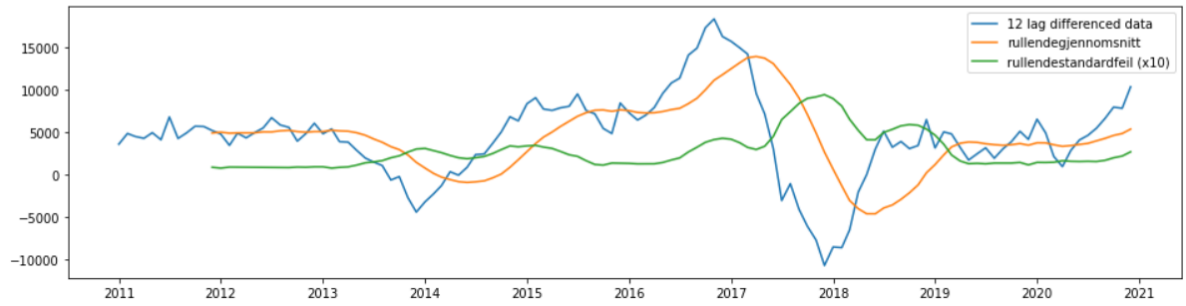
Første differensiering

```
 > Is the 1 lag differenced data stationary ?
Test statistic = -3.189
P-value = 0.021
Critical values :
        1%: -3.487517288664615 - The data is not stationary with 99% confidence
        5%: -2.8865777180380032 - The data is  stationary with 95% confidence
        10%: -2.5801239192052012 - The data is  stationary with 90% confidence
```



Sesong differensiering

```
 > Is the 12 lag differenced data stationary ?
Test statistic = -3.820
P-value = 0.003
Critical values :
        1%: -3.489057523907491 - The data is  stationary with 99% confidence
        5%: -2.887246327182993 - The data is  stationary with 95% confidence
        10%: -2.5804808802708528 - The data is  stationary with 90% confidence
```



9.7 SARIMA-prognoseresultater

Dato	Predikert gjennomsnitt	Nedre grense	Øvre grense
01.01.2021	91879.9278	88972.6124	94787.2431
01.02.2021	92940.7381	88473.1146	97408.3615
01.03.2021	92486.3228	86653.9082	98318.7374
01.04.2021	92826.6918	85747.457	99905.9267
01.05.2021	94381.1195	86146.7027	102615.536
01.06.2021	95954.7592	86642.6336	105266.885
01.07.2021	99472.5927	89150.2771	109794.908
01.08.2021	98467.3417	87194.3827	109740.301
01.09.2021	96870.8646	84700.0636	109041.666
01.10.2021	96718.3701	83696.7026	109740.038
01.11.2021	97236.3219	83405.6956	111066.948
01.12.2021	99256.7681	84654.6729	113858.863
01.01.2022	101152.318	85250.4809	117054.156
01.02.2022	102030.01	84822.4035	119237.616
01.03.2022	101440.741	82945.4703	119936.012
01.04.2022	101681.801	81930.3443	121433.258
01.05.2022	103163.096	82193.5967	124132.594
01.06.2022	104682.878	82536.083	126829.674

Tabell 9. 2 SARIMA-prognoseresultater

Norges miljø- og biovitenskapelige universitet
Noregs miljø- og biovitskapelege universitet
Norwegian University of Life Sciences

Postboks 5003
NO-1432 Ås
Norway